UNIVERSIDAD AUTÓNOMA DE MADRID

Departamento de Biología Molecular
Facultad de Ciencias

PROTEOGENÓMICA Y SPLICING ALTERNATIVO

TESIS DOCTORAL

Iakes Ezkurdia Garmendia

Madrid, 2015

UNIVERSIDAD AUTÓNOMA DE MADRID

Departamento de Biología Molecular
Facultad de Ciencias

PROTEOGENÓMICA Y SPLICING ALTERNATIVO

TESIS DOCTORAL

Memoria presentada para optar al grado de doctor en Ciencias por:

Iakes Ezkurdia Garmendia

Centro Nacional de Investigaciones Oncológicas (CNIO)

Madrid, 2015

Dirigida por: Dr. Michael Tress

D. Michael Tress, Investigador Principal del Programa de Biología Estructural y Biocomputación,

CERTIFICA: Que D. Iakes Ezkurdia Garmendia ha realizado el trabajo original de investigación "Proteogenómica y splicing alternativo", bajo su dirección en el Centro de Nacional de Investigaciones Oncológicas, para la obtención de Grado de Doctor.

Que considera que dicho trabajo reúne las condiciones necesarias para su presentación y defensa entre tribunal en la Facultad de Ciencias de la Universidad Autónoma de Madrid.

Y para que así conste, a los efectos oportunos, firma el presente certificado.

Madrid, a 2 de Noviembre de 2015

Dr. Michael Tress

“Why genes in pieces?”

Walter Gilbert

Resumen

La anotación manual de los genes codificantes de proteína requiere diversas fuentes de evidencia. Conseguir evidencia experimental de la expresión de las proteínas sigue siendo un reto técnico complicado. La mayoría de métodos se basan en predicciones computacionales y evidencia experimental a nivel de transcrito. La tecnología de espectrometría de masas ha avanzado considerablemente en las dos últimas décadas, situándola como una herramienta puntera para proyectos de anotación genómica. La espectrometría de masas permite la depuración y validación de genes codificantes y transcritos alternativos, así como la detección de nuevas regiones codificantes. La proteogenómica, una disciplina entre la genómica y la proteómica, requiere el desarrollo de métodos y estrategias computacionales para el análisis de datos a gran escala.

El objetivo principal de esta tesis es desarrollar métodos computacionales para el proceso y análisis de datos proteómicos y genómicos. Para ello se han diseñado varias estrategias de análisis de datos proteómicos a gran escala.

En la primera parte se aplican los flujos de trabajo diseñado para la búsqueda, validación y curación de resultados proteómicos, a partir de diversas fuentes de datos genómicos. La caracterización de isoformas alternativas y eventos de *splicing* en humano y ratón muestra tres grupos sobrerrepresentados. En concreto, las ribonucleoproteínas nucleares, las isoformas alternativas generadas a partir de exones homólogos, y las creadas a partir de deleciones pequeñas. El estudio se amplía utilizando una base de datos experimentales proteómicos mayor, y con ello se corrobora que la mayoría de genes expresa una proteína dominante. Se demuestra que los eventos de *splicing* detectados a nivel de proteína conservan los dominios funcionales. Finalmente, se ratifica que más del 20% de las isoformas de *splicing* están generadas por exones homólogos, que estas son específicas de tejido, y que están notablemente conservadas, advirtiéndose su posible relevancia a nivel celular.

En la última parte se utilizan los péptidos de ocho experimentos proteómicos a gran escala para caracterizar la isoforma más expresada del gen. La comparativa de la isoforma proteómica más expresada coincide con la de dos métodos ortólogos analizados. Uno basado en la conservación de función y estructura, y el otro basado en anotaciones genómicas corregidas por expertos. Los resultados muestran la tendencia hacia la expresión de una sola isoforma, independientemente del tejido, y confirman la idoneidad de APPRIS para la predicción de isoformas principales.

Abstract

The manual annotation of protein-coding genes is based on many diverse sources of evidence. Most support comes from computational predictions, genomic evidence and experimental expression at transcript level. Finding experimental evidence for the expression of proteins remains a difficult technical challenge, but mass spectrometry technology has advanced considerably in the past two decades, becoming an important tool for genomic annotation projects. Mass spectrometry also enables the refining and validation coding genes and alternative transcripts and detection of novel coding regions.

Proteogenomics, a discipline that unites genomics and proteomics requires the development of computational methods and strategies for data analysis on a large scale. The main objective of this thesis was to develop computational methods for processing and analyzing genomic and proteomic data. Several strategies to analyze large-scale proteomic data have been designed to achieve this goal.

In the first part workflows designed to search, validate and curate results from a variety of sources of proteomic data were applied as part of a pilot study. The characterization of alternative splice isoforms in human and mouse experiments highlighted three over-represented groups; specifically, ribonucleoproteins, alternative isoforms generated from homologous exons and those generated from small indels. The pilot study was later extended using a larger experimental proteomic data set. This second analysis confirmed that most genes express a dominant protein and demonstrated that splicing events detected at the protein level rarely break conserved functional domains. The large-scale study confirmed that more than 20% of splice isoforms are generated from homologous exons.
Many of these alternative homologous exons are tissue specific and all are remarkably conserved, highlighting their relevance at the cellular level.

Finally peptides from eight large-scale proteomic experiments are used to characterize a main experimental isoform. This main proteomics isoform matches those selected by two orthogonal methods, one predicted from conservation and protein functional and structure features, and the other annotated by manual annotators based on genomic evidence. The results show clearly that almost all genes have a principal protein isoform regardless of tissue.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Introducción

1.1. *Splicing* alternativo

En 1978, Walter Gilbert (Gilbert, 1978) contradijo el dogma fundamental de un gen, una cadena polipeptídica proponiendo que el uso diferente de los exones de un mismo gen (el *splicing* alternativo) podría dar lugar a proteínas diferentes.

El *splicing* es un mecanismo regulatorio a través del cual los exones se combinan para dar lugar a distintos transcritos de ARN mensajero (mRNA, en inglés messenger RNA). Si el proceso de *splicing* puede generar transcritos distintos a partir de la combinación de exones se denomina *splicing* alternativo. Estos transcritos podrían ser traducidos en diferentes proteínas derivadas de un mismo gen (Black, 2003a). Un ejemplo de la diversidad a la que podría dar lugar este mecanismo es el caso del gen *Dscam* en *Drosophila* que tiene el potencial de codificar 38016 variantes de *splicing*, el equivalente a tres veces el número total de genes en *Drosophila* (Black, 2000). Experimentos de transcriptómica a gran escala han demostrado la importancia del *splicing* alternativo (Harrow *et al.*, 2006; Xue *et al.*, 2009).

1.1.1. El mecanismo de *splicing*

Los genes en eucariotas superiores están compuestos por exones e intrones. El proceso de splicing elimina los intrones del transcrito primario (pre-mRNA, del inglés precursor mRNA), dejando únicamente los exones en el transcrito maduro (S E Leff & Rosenfeld, 2003). El espliceosoma, encargado del mecanismo de *splicing*, es un gran complejo molecular formado por varios RNA nucleares pequeños (snRNA, del inglés small nuclear RNA) y más de 150 proteínas (Will & Lührmann, 2011). Los snRNA están unidos a varias proteínas, formando ribonucleoproteínas nucleares pequeñas (snRNPs). De éstas, seis snRNPs (llamadas U1, U2, U3, U4, U5 y U6) se encargan del proceso de *splicing* y son responsables de la actividad catalítica principal del espliceosoma (Will & Lührmann, 2011).

En el reconocimiento de un intrón en la secuencia pre-mRNA, participan tres señales de *splicing*: el sitio 5' donador de *splicing*, el sitio 3' aceptor de *splicing* y la secuencia de ramificación situada a unos 40 nucleótidos de extremo 3' del intrón (Wu & Krainer, 1999). El sitio donante tiene una secuencia consenso igual a GTRAGT (donde R puede ser A o G) en mamíferos. En el 5' terminal del intrón está presente de forma invariable el dinucleótido AG al que le preceden una serie de nucleótidos de pirimidina llamada tracto de polipirimidina. El punto de ramificación adenosina contiene la secuencia consenso YTRA (donde Y puede ser C o T).

Durante el *splicing* de un intrón, la unidad snRNP U1 comienza formando un complejo en el sitio donador de *splicing* reconociendo pares de bases específicas entre el snRNA y el mRNA (Nagai *et al.*, 2001). Posteriormente, la unidad U2AF se une con el tracto de polipirimidina y el sitio aceptor. Las unidades U4, U5 y U6 hacen de nexo entre los dos puntos, haciendo que el intrón se pliegue, el mRNA se rompe en el sitio donador y el extremo 5' del intrón se une a la secuencia de ramificación adenosina formando un lazo. Esto acerca la región del exón permitiendo la unión en el sitio donde se corta el aceptor y el intrón es liberado (Cheng & Menees, 2011).

1.1.2. Eventos de *splicing* alternativo

Los eventos de *splicing* se pueden clasificar a partir de tres mecanismos fundamentales: selección de promotores alternativos, selección de sitios de poliadenilación alternativos y *splicing* alternativo por selección de exones (Black, 2003b; Matlin *et al.*, 2005; Pan *et al.*, 2008; Sammeth *et al.*, 2008). La selección de promotores tiene lugar cuando se utilizan dos promotores alternativos, dando lugar a dos mRNAs diferentes (Figura 1). La selección de sitios de poliadenilación alternativa ocurre cuando existe más de un sitio de unión en la cola poli-A dando lugar a un juego de exones diferentes. Este mecanismo es utilizado para la producción de anticuerpos dando lugar a dominios C-terminales alternativos. La elección del sitio de unión del promotor o del sitio de unión de la cola se hace en base a factores de transcripción específicos del tipo de célula.

La selección de exones es el mecanismo de *splicing* más común para expresar diferentes proteínas. Este mecanismo requiere la selección de diferentes sitios de *splicing* y da lugar a la inclusión o exclusión de exones completos (Sammeth *et al.*, 2008) (Figura 1). Los procesos de *splicing* pueden regularse en función del tipo de célula o como respuesta a señales relacionadas con el desarrollo o el entorno. En cualquier caso, todavía no se conocen todos los factores de los mecanismos involucrados en el reconocimiento de los sitios de *splicing* y no se puede asegurar que los procesos de reconocimiento de *splicing* funcionen en el 100% de los casos (Sorek *et al.*, 2004).

El mecanismo de trans-*splicing* es un mecanismo especial y poco frecuente que da lugar a la unión de segmentos a partir de dos transcritos primarios diferentes (Iwasaki *et al.*, 2009). Ocurre con más frecuencia en trypanosomas, nematodos y cloroplastos de células de plantas.

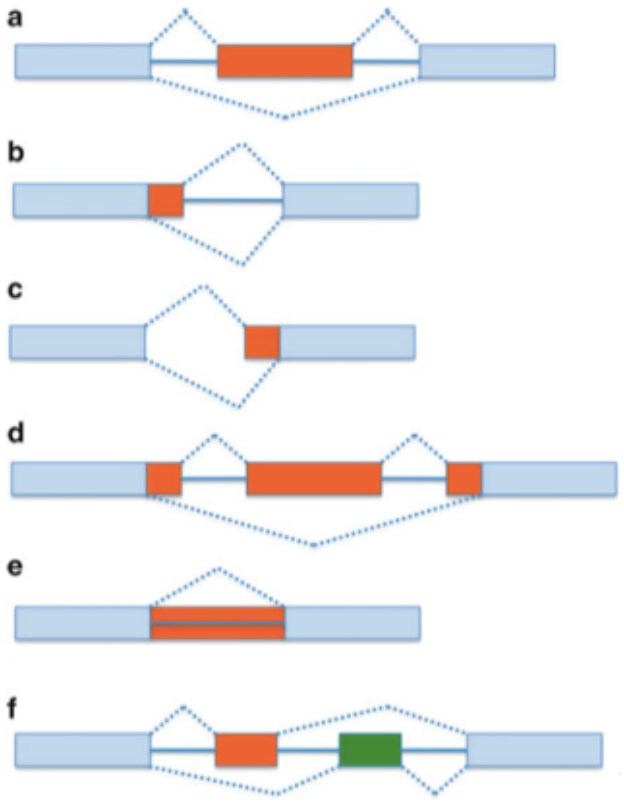


Figura 1. Eventos de *splicing* alternativo, (a) Inclusión o exclusión de uno o más exones (*exon skipping*), (b) sitio donador alternativo, (c) sitio aceptor alternativo, (d) patrones de *splicing* complejos, (e) retención de intrones, (f) exones mutuamente excluyentes.

1.1.3. Regulación de *splicing* alternativo

Los sitios de *splicing* y puntos de ramificación en organismos superiores no son del todo específicos y tienden a confundirse con muchas secuencias similares. Además, el *splicing* depende del tipo de tejido celular, el estado de desarrollo o los estímulos externos (David & Manley, 2008). Estos factores sugieren que existen señales externas involucradas en los procesos de regulación y *splicing*. Los motivos localizados en los exones e intrones, también denominados reguladores CIS, contribuyen a regular la frecuencia con la que se realizan los procesos transcripcionales potenciándolos o silenciándolos; se suelen denominar ESE (del inglés, *exonic splicing enhancer*), ESS (del inglés, *exonic splicing silencer*), ISE (del inglés, *intronic splicing enhancer*) y ISS (del inglés, *intronic splicing silencer*) (Wang & Burge, 2008). Los elementos CIS y TRANS, controlan la estabilidad del mRNA, regulando las cantidades de expresión. La mayoría de las secuencias regulatorias potenciadoras, se unen mediante algunas de las 6 proteínas con dominios ricos en serina y arginina llamadas proteínas SR (el símbolo de los residuos de serina y arginina) (Long & Caceres, 2009). Los silenciadores pueden variar en secuencia, y se unen normalmente a ribonucleoproteínas heterogéneas-nucleares (hnRNPs). Algunos motivos pueden comportarse como potenciadores o silenciadores dependiendo de su distancia al sitio de *splicing*. El *splicing* constitutivo y alternativo está también regulado a partir de señales de *splicing* adicionales y factores de *splicing* que permiten distinguir sitios de *splicing* ambiguos.

Para entender bien como se regula el *splicing* alternativo, es necesario conocer a fondo los motivos y factores constitutivos de *splicing*. Los mecanismos de regulación, funcionan muchas veces combinando diferentes factores de *splicing* y mecanismos de corrección que son controlados en ciertos tejidos o como respuesta a estímulos específicos (Barash & Barash, 2010). Se han estudiado y caracterizado procesos de regulación específicos para el control de *splicing* alternativo como por ejemplo, la determinación del sexo en *Drosophila*

(McIntyre *et al.*, 2006) y la transmisión sináptica en mamíferos (Ule *et al.*, 2005). La variación de concentración relativa de factores generales de *splicing* dependientes del entorno puede influir en el *splicing* de muchos transcritos. Muchos factores de *splicing* se autoregulan también a partir del *splicing* alternativo de su mismo mRNA. Los estados de fosforilación de estos factores de *splicing*, cambian su localización subcelular impidiendo que afecten al *splicing*.

Existen además otros factores que afectan al *splicing* como la velocidad de transcripción del gen que puede favorecer la inclusión de exones en el transcrito maduro cuando la polimerasa se ralentiza durante la elongación, o la estructura secundaria de los mRNA (Shepard & Hertel, 2008).

La estructura de la cromatina también interviene en el *splicing* alternativo. Estudios en humano, han encontrado correlaciones entre la posición de los nucleosomas y los niveles de inclusión de exones (Tilgner *et al.*, 2009). Los exones constitutivos aparecen frecuentemente incluidos en el nucleosoma, mientras que la frecuencia de inclusión de los exones alternativos es muy baja y la de los intrones la menos frecuente de todas. Algunas modificaciones epigenéticas funcionan a veces como sitios de unión de *splicing* favoreciendo la inclusión de exones. El nucleosoma añade un nivel de complejidad a la regulación de *splicing* en eucariotas.

1.1.4. Métodos de asociación de genoma completo

Los transcritos alternativos se pueden encontrar a partir de marcadores de secuencia expresada (ESTs) (Modrek *et al.*, 2001a), microarrays de ácido desoxirribonuclico (DNA, del ingés deoxyribonucleic acid) (Johnson *et al.*, 2003), secuenciación exhaustiva (*deep sequencing*) (Bryant *et al.*, 2012), entrecruzamiento e inmunoprecipitación (CLIP, del inglés cross-linking immunoprecipitation) (David & Manley, 2008) y perfiles ribosomales (en inglés, *ribosomal profiling*).

Las muestras de ESTs (en la actualidad, alrededor de 8 millones en humanos) permiten la detección de eventos de *splicing* alternativos en genomas completos. Los ESTs y los mRNA son alineados contra el genoma. Las zonas alineadas corresponden generalmente a los exones donde y las zonas de gaps a los intrones. Debido a la ambigüedad del alineamiento de empalmes exón-exón, se suelen emplear algoritmos específicos para los cálculos (Yeo *et al.*, 2005). Los eventos de *splicing* alternativo se identifican a partir de los exones que aparecen presentes en algunos ESTs y los que están ausentes en otros. Eventos como la retención de intrones, así como otros eventos de *splicing* se detectan a partir de estos alineamientos. Los ESTs también se emplean para la predicción de variantes de *splicing* específicas para ciertos tejidos o enfermedades como el cáncer. Los inconvenientes de esta tecnología son los altos costes, la contaminación genómica, el sesgo en el clonado, confusión con parálogos y una baja sensibilidad para los transcritos poco abundantes.

En el análisis de *exon array* se utilizan arrays de fragmentos de DNA que representan exones individuales o de zonas de unión (Sugnet *et al.*, 2006). El *array* se sondea con DNA complementario (cDNA, del inglés complementary DNA) marcado de los tejidos de interés. La sonda de cDNA se une al DNA de los exones que están incluidos en el mRNA del tejido seleccionado, o del sitio de unión del *splicing*. La intensidad de las sondas se utiliza para determinar que eventos de *splicing* están presentes y se calculan los ratios para medir la presencia de las variantes de *splicing* de mRNA.

Estudios basados en microarrays han mostrado que el 74% de los genes sufren *splicing* alternativo (Johnson *et al.*, 2003). Los microarrays se pueden utilizar para detectar y analizar eventos de *splicing* específicos de tejido, para estudiar la regulación y para encontrar relaciones entre enfermedad y *splicing* alternativo. Las limitaciones de esta tecnología están supeditadas a la cobertura de las sondas, a las alteraciones de la hibridación cruzada y a los análisis complejos requeridos.

Las tecnologías de secuenciación exhaustiva (RNAseq) también se utilizan para estudiar las asociaciones de genoma completo (GWAS, en inglés Genome-Wide Association Study) de variación de *splicing* (He *et al.*, 2009). El *splicing* alternativo se detecta alineando las lecturas

de RNAseq y las uniones de *splicing*. La fiabilidad de las predicciones de uniones de *splicing* esta determinada por el número de lecturas que mapean al sitio de unión, el número de errores para cada lectura, las distancias en los mapeos al sitio de unión, y los errores en las zonas de unión. A pesar del potencial de esta tecnología, el análisis de los datos de RNAseq sigue siendo un reto importante (Hayer *et al.*, 2015).

El entrecruzamiento e inmunoprecipitación une las proteínas a las moléculas de RNA en un tejido durante el *splicing* utilizando radiación UV. Se utilizan anticuerpos específicos haciendo que las proteínas reguladoras que actúan en "TRANS" precipiten. Después, el RNA unido a esa proteína se aísla y se clona, mostrando las secuencias diana de la proteína (Ule *et al.*, 2006).

1.1.5. Las implicaciones del *splicing* alternativo

Las isoformas producidas por el *splicing* alternativo pueden tener características diferentes con respecto a la isoforma de referencia como la composición de dominios, la afinidad de unión a ligando, la localización subcelular, y la vida media (Tress *et al.*, 2007b; Light & Elofsson, 2013). La inserción o deleción de dominios, hélices transmembrana o péptidos señal pueden dar lugar a modificaciones en la función o localización de la proteína. Las isoformas transmembrana que pierden los exones que codifican las hélices transmembrana pueden generar isoformas solubles (Tress *et al.*, 2007b). Hay un ligero sesgo hacia la producción de isoformas que pierden unidades funcionales completas como los dominios de proteínas (Kriventseva *et al.*, 2003a). Se ha constatado que la regulación del *splicing* alternativo afecta a diferentes procesos biológicos como el desarrollo de la sinapsis, la orientación de exones, y la activación de linfocitos T . Sin embargo, dado que el *splicing* alternativo ocurre en la mayoría de los genes humanos que contienen múltiples exones, es posible que el *splicing* alternativo esté presente también en la mayoría de procesos biológicos.

Para evitar la generación de proteínas anormales debido a un procesamiento incorrecto del mRNA, las células han desarrollado mecanismos de vigilancia encargados de degradar

algunos transcritos. La degradación del RNA mensajero mediada por mutaciones terminadoras o NMD (Nonsense mediated decay) (Baker & Parker, 2004) es un mecanismo que degrada los mRNA para los que se detecta un codón de terminación prematuro. El mecanismo de NMD detecta esos codones de terminación que han sido introducidos debido a la alteración de un marco abierto de lectura en el 3' terminal, muchas veces como resultado de un salto de exón en el *splicing*.

Estos errores de *splicing* son la causa de un número de enfermedades en humano que han sido muy bien caracterizadas, como la fibrosis quística o el síndrome de Hutchinson–Gilford. Los cambios que ocurren en la maquinaria de *splicing* pueden dar lugar a errores de *splicing* en múltiples transcritos. Las alteraciones en los nucleótidos únicos también pueden darse en los sitios de *splicing* o en elementos CIS provocando errores en los mRNA. Se ha estimado que la mitad de las enfermedades humanas provocadas por mutaciones están relacionadas con mecanismos de *splicing* (López-Bigas *et al.*, 2005). Las proteínas codificadas por variantes de *splicing* que están asociadas a diferentes enfermedades, podrían convertirse en dianas terapéuticas.

1.1.6. Evolución del *splicing* alternativo

El *splicing* alternativo está presente en el árbol genealógico de las especies eucariotas y tiene el potencial para incrementar la diversidad del transcriptoma y el proteoma de estas especies. Parece que es más abundante en eucariotas superiores que en eucariotas inferiores, con un número mayor de genes en los vertebrados. Las plantas tienen un nivel más bajo de *splicing* alternativo. También se da en algunos casos en bacteria y arqueas.

Estudios recientes donde se comparan especies diferentes han servido para comprender mejor cómo se genera y se reconoce la arquitectura de los intrones y exones. La arquitectura del *splicing* alternativo varía entre eucariotas. Existen restricciones en el reconocimiento del *splicing* relacionadas con la longitud de los intrones y exones, con intrones largos y exones cortos en vertebrados e intrones cortos y exones largos en eucariotas (Hawkin, 1988).

Los exones susceptibles de *splicing* alternativo se pueden generar a partir de tres mecanismos evolutivos: barajado de exones (exon shuffling), conversión de intrones y transición de exones constitutivos (Keren *et al.*, 2010). El proceso de barajado de exones ocurre cuando un exón es insertado en un gen o es duplicado dentro del mismo gen. La conversión de un intrón a exón ocurre cuando una sección genómica se convierte *de novo* en un exón. Los mecanismos de transición implican la conversión de un exón constitutivo en un exón alternativo.

Existen dos hipótesis contrapuestas que explicarían el origen biológico de los intrones. En la hipótesis de *introns early* (Penny *et al.*, 2009) se asume que surgió un mecanismo que permitiría eliminar las regiones del mensajero primario que contenían codones de parada, para permitir proteínas de longitudes mayores y con más diversidad funcional. Los marcos de lectura, serían los exones de hoy en día, mientras que los que los intrones serían las juntas de *splicing* que cumplían la función de eliminar las señales de parada. En la hipótesis de *introns late*, se asume que el origen de los genes estaría en pequeños marcos de lectura que habrían evolucionado a partir de mecanismos de duplicación y fusión. Los intrones serían el resultado de inserciones de DNA foráneo en los genes. La degeneración de las señales de *splicing* ancestrales y que el spliceosoma estuviera presente en los eucariotas inferiores sugiere que el *splicing* alternativo eucariota surgió en las primeras etapas de la evolución. Por ejemplo los mecanismos de NMD están presentes en animales, fungi y plantas, y se han encontrado similitudes en los patrones de *splicing* entre genomas de especies muy lejanas (Lareau & Brenner, 2015).

1.1.7. Bioinformática y *splicing* alternativo

La bioinformática ha tenido un gran impacto en el campo del *splicing* alternativo con la llegada de datos generados a partir de EST, cDNA y diferentes proyectos de secuenciación. En los estudios de GWAS en *splicing* alternativo que se han llevado a cabo, se estima que el *splicing* alternativo se da en un 34-94% de los genes un porcentaje que aumenta en función de en estudios más recientes (Pan *et al.*, 2008; Wang *et al.*, 2008). Estos datos de *splicing*

alternativo han sido analizados con el propósito de caracterizar el impacto funcional de las variantes de *splicing* a partir de distintas aproximaciones (Blencowe & Graveley, 2008).

Muchas de las herramientas bioinformáticas disponibles están relacionadas con la detección de *splicing* alternativo a partir de la correlación de secuencias EST con mRNA y secuencias genómicas. La calidad de las predicciónes de *splicing* alternativo están supeditadas a los diferentes tipos de *splicing* alternativo. Los sitios donadores y aceptores alternativos así como los sitios de salto de exones son fáciles de detectar e interpretar, mientras que los sitios de retención de intrones son generalmente lo más difícil de predecir debido a las alteraciones experimentales. Se utilizan algoritmos de secuenciación rápida para mapear todos los mRNA y los ESTs contra el genoma. Los algoritmos de reconocimiento tratan de distinguir los fragmentos procedentes de transcritos alternativos de las alteraciones producidas por la limitación la cobertura de los EST, los errores de ensamblaje genómicos, la contaminación genómica, errores de orientación de los EST, errores de secuenciación, o fragmentación de EST (Eyras, 2004).

Las secuencias genómicas aportan una información indispensable para la detección y validación de eventos de splicing a nivel de transcrito, ya que gran parte de la información se deriva de los intrones. La secuencia genómica también ayuda a resolver problemas de agrupamiento de ESTs y a distinguir parálogos que pudieran interferir en la calidad de las predicciones. Los métodos que no están basados en ESTs, pueden utilizar características como la conservación, las regiones limítrofes de los intrones, las longitudes de los intrones y exones, y la composición de nucleótidos para predecir exones (Zhang, 2002).

Otro papel de la bioinformática en el *splicing* alternativo consiste en la construcción de modelos genéticos alternativos integrando los datos de los exones para generar isoformas plausibles.

Algunos de los algoritmos bioinformáticos relacionados con la caracterización del *splicing* alternativo se están desarrollando entorno a proyectos genómicos de gran escala como ENCODE (Anon 2004). Estos algoritmos se están empezando a utilizar en la anotación

automática y la predicción del espliceosoma. GENCODE (Harrow *et al.*, 2006) es un subproyecto de ENCODE que tiene como objetivo principal la identificación de todos los genes codificantes, los no codificantes y de pseudogenes. La validación y mejora de estos algoritmos, se lleva a cabo contrastándolos con anotaciones a mano de gran calidad.

1.2. La complejidad de los proteomas

El proteoma celular ha sido definido como el total de las proteínas expresadas por una célula particular bajo unas condiciones específicas (Wilkins *et al.*, 1996). El número de proteínas en un proteoma eucariota depende de al menos tres mecanismos celulares: la regulación de la expresión génica, los eventos post-transcripcionales y las modificaciones post-transduccionales. Para determinar la composición del proteoma celular, mediante por ejemplo, la espectrometría de masas, se requiere un conocimiento profundo de estos tres mecanismos sin los cuales no podemos definirla *a priori*. La expresión génica es sólo el primero de lo niveles de la complejidad del proteoma. La identificación precisa en los genes eucariotas es técnicamente difícil, aún no hay un método que sea capaz de lograrlo de manera exhaustiva y los algoritmos de predicción siguen mejorándose (Zhang & Zhang, 2002; Wang *et al.*, 2003; Brent & Michael, 2005).

Como ya hemos comentado antes, los genes multi-exónicos de los organismos eucariotas, pueden producir diferentes transcritos mRNA a partir de mecanismos del *splicing* alternativo del pre-RNA. Existen además más capas de complejidad como la de los polimorfismos de nucleótido simple (SNP) que pueden dar lugar a la mutación de aminoácidos en la proteína resultante. Los SNPs son el tipo más común de variación genética entre poblaciones de individuos. Los SNPs no-sinónimos pueden también influir en las modificaciones post-transduccionales de las proteínas (PTM, del inglés Post Translational Modifications) como la fosforilación (Ryu *et al.*, 2009). Recientemente se han publicado estudios a gran escala para analizar polimorfismos en *Saccharomyces cerevisiae* (Schacherer *et al.*, 2009), *Arabidopsis thaliana* (Clark *et al.*, 2007), y *Mus musculus* (Frazer *et al.*, 2007). En los resultados iniciales del "Proyecto de los 1000 genomas", un proyecto con el objetivo de caracterizar sistemáticamente la variación genética humana (1000 Genomes Project

Consortium, 2010), se han identificado 15 millones de SNPs. Entre estos, más de 1 millón son *indels* y hay mas de 20,000 variantes estructurales, el 55% de los SNPs encontrados no habían sido descritos anteriormente. Existe además un nivel adicional de complejidad formado por las múltiples variantes proteicas derivadas de las PTMs. Las PTMs son eventos enzimáticos que ocurren durante o después de la síntesis de proteínas y están causados por modificaciones químicas de uno o varios aminoácidos. Son un componente importante del sistema de señalización celular. Entre los ejemplos frecuentes de PTMs se incluyen fosforilaciones, glicosilaciones, metilaciones, acetilaciones, ubiquitinizaciones y lipidaciones (Aebersold & Mann, 2003). Hoy en día se conocen más de 200 tipos diferentes de PTMs y las estimaciones más conservadoras sugieren que una proteína podría ser modificada por unos 10 PTMs diferentes (Cox & Mann, 2011). Si tenemos en cuenta sólo la fosforilación, uno de las PTMs mejor estudiados, existen más de 500,000 sitios de fosforilación predichos en el proteoma humano.

A la complejidad del proteoma celular, hay que añadir su naturaleza dinámica. La célula modifica el proteoma en función a la respuesta a diversos estímulos e infinidad de factores ambientales que dificultan la predicción del mismo y complican su estudio. Con los nuevos avances tecnológicos se ha incrementando la fracción del proteoma que se puede medir, sin embargo la naturaleza dinámica del proteoma, impide dar por finalizada la identificación de proteínas. Esto implica que actualmente, se podrían considerar incompletos todos los estudios de proteómica (Ahrens *et al.*, 2010).

Los enormes avances técnicos en proteómica de la última década han posibilitado la detección de una fracción importante de proteínas en diferentes organismos y líneas celulares, que implica una inversión considerable en tiempo, instrumentación especializada y capacidad de computación. No obstante, la detección de todas las proteínas de un organismo sigue siendo un reto, tanto por la incapacidad de predecir el proteoma de cualquier célula (Ahrens *et al.*, 2010), como por las limitaciones técnicas (Omenn *et al.*, 2006). Se estima que la digestión con tripsina de un proteoma puede producir hasta 1 millón de péptidos diferentes y la complejidad de la muestra puede verse aumentada artificialmente en función de los protocolos que se utilicen antes del análisis de espectrometría de masas.

Las limitaciones en la velocidad de escaneo de los espectrómetros de masas impide la fragmentación secuencial de un número muy grande de péptidos diferentes, como los obtenidos a partir de la digestión de proteomas celulares completos. Esto hace que los análisis masivos de péptidos trípticos (*shotgun*) sean incompletos cubriendo solo una parte del proteoma que tiende a ser la de los péptidos más abundantes (Abu-Farha *et al.*, 2009; Gstaiger *et al.*, 2009). Para atajar este problema, se suelen emplear estrategias de fraccionamiento y separación de proteomas complejos. En dos estudios publicados recientemente se ha investigado el proteoma de dos líneas celulares de mamíferos a partir de espectrometría de masas. En estos estudios se han identificado más de 10,000 proteínas que cubren una proporción significativa del los proteomas de estas líneas celulares (Beck *et al.*, 2011; Nagaraj *et al.*, 2011). En la Tabla 1 listamos varios organismos para los que sus proteomas se han investigado en profundidad. Se muestra la proporción de proteínas identificadas y los ORFs predichos en diferentes proteomas y como se puede observar, la cobertura del proteoma esta relacionada con la complejidad del organismo.

1.2.1. Polimorfismos de nucleótido simple

Las limitaciones de las tecnologías actuales para la investigación en proteómica se hacen patentes cuando tenemos en cuenta el número SNPs. La identificación de SNPs a partir de algoritmos de búsqueda en bases de datos implica una alta tasa de falsos positivos. La detección de polimorfismos por proteómica *shotgun* ha sido estudiada por varios grupos. Las estrategias incluyen ampliar la base de datos de búsqueda con los SNPs conocidos (Gatlin *et al.*, 2000) y el filtrado de espectros de alta calidad junto con la búsqueda de espectros sin asignar (Nesvizhskii *et al.*, 2006). Bunger et al. (Bunger *et al.*, 2007) utilizan un algoritmo *ad hoc* de predicción de péptidos y una base de datos señuelo generada a partir de SNPs aleatorios para encontrar nsSNP. Los autores identifican un total de 629 nsSNPs a partir de datos de shotgun a partir de muestras fraccionadas de líneas celulares humanas de cáncer de mama.

1.2.2. Modificaciones postraduccionales

Las mismas limitaciones existentes para la detección de SNPs y variantes de *splicing* existen en la caracterización de PTMs en proteomas completos. Los PTMs son a menudo subestequiométricos y por tanto, difíciles de detectar. Los análisis clásicos de PTMs por espectrometría de masas, requieren varios pasos: primero se enriquece el subproteoma a partir de péptidos que incluyen las modificaciones a estudiar, segundo se detectan e identifican los péptidos modificados a partir de la digestión proteolítica y tercero se identifican los residuos modificados de la secuencia peptídica. Los avances en la instrumentación y los protocolos de enriquecimiento han permitido la identificación de miles de PTMs como las fosforilaciones, acetilaciones, y metilaciones en diferentes organismos y tejidos. Hoy en día, existen repositorios públicos de datos como PhosphositePlus (Hornbeck *et al.*, 2004), Phosida (Gnad *et al.*, 2011), (Bodenmiller *et al.*, 2007) o Uniprot (www.uniprot.org) que contienen este tipo de información. La base de datos PhosphositePlus, por ejemplo, integra datos de experimentos de alto y bajo rendimiento y contiene más de 1,000,000 de sitios de fosforilación identificados en diversos organismos. La enorme cantidad de datos acumulada ha facilitado el desarrollo de varias herramientas de predicción de PTMs (Eisenhaber & Eisenhaber, 2010). Algunos de los más conocidos son Scansite (Obenauer *et al.*, 2003) que predice sitios de fosforilación a partir de la búsqueda de patrones correspondientes a motivos de quinasas consenso en bases de datos de proteinas, Phosida (Gnad *et al.*, 2011) que ofrece herramientas de predicción para la fosforilación y acetilación de lisina, y UbPred (Radivojac *et al.*, 2010), que identifica sitios de ubiquitinación. El problema general de los predictores de sitios PTM, es la elevada tasa de falsa detección (FDR, del inglés False Discovery Rate) que impide hacer predicciones precisas del alcance de los PTMs en la célula.

La complejidad de los proteomas surge de los mecanismos celulares de la expresión de proteínas y de los mecanismos de regulación. Muchas veces las mejoras en la instrumentación, en los protocolos de preparación de muestras y en los flujos de análisis de datos, permiten la detección de un número cada vez más grande de proteínas. En este contexto, las aproximaciones para encontrar proteínas nuevas de manera rápida y precisa son

cada vez más importantes ya que servirán a los biólogos para la caracterización de las funciones biológicas.

Organismo	Proteínas identificadas en experimentos de proteómica	Genes codificantes anotados	Cobertura del proteoma	Transcritos de proteína anotados	Genes codificantes predichos	nsSNP	Fuente de anotación genómica
Bacteria y arquea							
Mycoplasma pneumoniae M129	620	688	0.9	-	779	-	CMR
Thermoplasma acidophilum	1025	1478	0.69	-	1630	-	CMR
Staphylococcus aureus COL	1703	2681	0.63	-	2716	-	CMR
Leptospira interrogans serovar Copenhageni	2221	3660	0.61	-	4475	-	CMR
Eukariotas							
Saccharomyces cerevisiae	4399	6696	0.66	7130	7940	64167	SGD rel. 64.10
Drosophila melanogaster	9263	13781	0.67	23017	19437	459154	Ensembl rel. BDGP 5.25
Pristionchus pacificus	4029	5211	0.77	5211	24217	-	Wormbase rel. W228
Caenorhabditis elegans	6779	23358	0.29	29872	25391	2493	Wormbase rel. W22
Arabidopsis thaliana	13029	27299	0.48	34183	23868	896537	Ensembl rel. 64.10
Mus musculus	7686	22234	0.35	44337	46375	30463	Ensembl rel. 64.37
Homo sapiens	12141	20996	0.58	72065	47019	128519	Ensembl rel. 64.3

Tabla 1. La cobertura del proteoma está calculada como el porcentaje de proteínas detectadas en experimentos de proteómica frente a las anotadas en todos los genes codificantes; nsSNP, polimorfismos no-sinónimos codificantes (Clarke *et al.*, 2012), CMR, *The Comprehensive Microbial Resource* (Peterson *et al.*, 2001); SGD, *Saccharomyces cerevisiae Genome Database* (Cherry *et al.*, 1997); Ensembl rel 63, The Ensembl Project (Hubbard *et al.*, 2002); Wormbase W228, (Harris *et al.*, 2010). Las predicciones de genes codificantes para bacteria están basados en el algoritmo Glimmer (Delcher *et al.*, 2007)y las anotaciones están extraidas de *GenBank*. La predicción de genes codificantes de proteína para *Pristionchus pacificus* y *Caenorhabditis elegans* están extraídas de *Wormpep* y están basadas en *GeneFinder* (P. Green, unpubl.). Para humano, ratón, *Drosophila melanogaster* y *Arabidopsis thaliana* los

genes codificantes anotados son aquellos que tienen al menos un transcrito con una secuencia en un repositorio externo a Ensembl para la misma especie. En la columna de transcritos anotados se muestra el número total de variantes de *splicing* confirmados experimentalmente. Los polimorfismos no-sinonimos codificantes están extraidos de la base de datos de variaciones EF3 *Saccharomyces cerevisiae*, BDGP 5 *Drosophila melanogaster*, ws220 *Caenorhabditis elegans*, TAIR10 *Arabidopsis thaliana*, dbSNP128 *Mus musculus* y dbSNP132 *Homo sapiens*. Para aquellos casos en los que estaban disponibles varias cepas, se han tenido en cuenta todas ellas.

1.3. *Splicing* alternativo y proteómica

Como se ha comentado, los experimentos de transcriptómica a gran escala sugieren que la mayoría de los genes codificantes podrían dar lugar a múltiples transcritos. Los análisis iniciales de RNAseq y ESTs predecían que podría haber unos 100,000 eventos de splicing detectables estimando que la mayoría de genes humanos con múltiples exones darían lugar a splicing alternativo (Mollet *et al.*, 2010). Sin embrago, todavía no hay un consenso claro sobre el número de transcritos que se podrían producir. Estudios recientes muestran una gran disparidad de resultados, el trabajo de Uhlen (Uhlen *et al.*, 2015) basado en datos de RNAseq a gran escala se sugiere que un 72% de los genes estarían expresando más de una isoforma mientras que en el de Hu (Hu *et al.*, 2015) predicen que habría 205,000 transcritos codificantes de proteína.

La abundancia de transcritos detectados ha nivel de gen ha dado lugar a suponer que si estos fueran traducidos en proteínas, estas podrían estar jugando un papel determinante en la complejidad de los mamíferos (Nilsen & Graveley, 2010). Cuantas de estas isoformas alternativas son realmente funcionales a nivel de proteína sigue siendo una pregunta sin resolver de gran importancia en la biología eucariota.

Para demostrar la traducción de estos transcritos alternativos se han utilizado diversas aproximaciones. En algunos casos, experimentos individuales han aportado evidencia de expresión a nivel de proteína para las isoformas de genes específicos (Lane *et al.*, 2014).

También se ha utilizado anticuerpos para la detección de isoformas alternativas (Uhlen *et al.*, 2015), sin embargo la falta de especificidad de la mayoría de anticuerpos hace que su discriminación sea casi imposible. Para poder aplicar esta tecnología satisfactoriamente los anticuerpos tendrían que haber sido diseñados para detectar un único evento de splicing. También se ha utilizado recientemente la técnica de perfil ribosomal (*ribosome profiling*, o RPF) para determinar zonas codificantes (Vanderperre *et al.*, 2013; Bazzini *et al.*, 2014), sin embargo estas técnicas presentan algunos problemas (Gerashchenko & Gladyshev, 2014). Además, esta tecnología depende de algoritmos de reconstrucción para la predicción de isoformas que todavía no son del todo fiables (Steijger *et al.*, 2013; Lahens *et al.*, 2014; Hayer *et al.*, 2015).

La mayor fuente de evidencia de splicing alternativo a nivel de proteína es la espectrometría de masas (Farrah *et al.*, 2012). La espectrometría de masas se ha convertido en una herramienta de gran valor para la validación de anotaciones genéticas (Tanner *et al.*, 2007a; Brosch *et al.*, 2011). El Human Proteome Project (Legrain *et al.*, 2011) tiene como objetivo la detección de al menos una proteína para cada uno de los genes codificantes anotados en el genoma humano.

Hasta ahora se ha encontrado poca evidencia fiable de isoformas alternativas en experimentos de proteómica. El grupo de Hubbard (Brosch *et al.*, 2011) detectó 53 genes en ratón que expresaban más de una isoforma alternativa buscando en una base de datos de gran tamaño. Se han utilizado aproximaciones similares para identificar isoformas alternativas en humano (Tanner *et al.*, 2007a), *Rattus norvegicus* (Low *et al.*, 2013), *Arabidopsis* (Castellana *et al.*, 2008), *Drosophila* (Tress *et al.*, 2008) y *Aspergillus flavus* (Chang *et al.*, 2010). Recientemente Ning y Nesvizhskii (Ning & Nesvizhskii, 2010) examinaron la viabilidad de la utilización de datos de espectrometría de masas (MS, del inglés Mass spectroscospy) para la identificación de isoformas de splicing alternativo no descritas a partir de búsquedas de espectros MS/MS, utilizando repositorios públicos de experimentos de proteómica para tejidos de ratón y bases de datos de RNAseq. Los autores demostraron la correlación entre la probabilidad de identificación de un péptido a partir de datos MS/MS y el número de *reads* en los datos de RNAseq para el mismo gen. Sin embargo, el número de péptidos nuevos que

fueron identificados por espectros MS/MS era sustancialmente menor al número esperado basado en predicciones *in silico*.

Recientemente se han publicado algunos estudios proteómicos en los que afirman haber encontrado muchas isoformas alternativas expresadas a nivel de proteína. Sin embargo, para los trabajos en los que pudimos verificar los resultados publicados, vimos que la abundancia de isoformas detectadas se debían a fallos de mapeo o a la subestimación de falsos positivos. Como se muestra en la sección de resultados de esta memoria, las isoformas alternativas a nivel de proteína solo pueden detectarse cuando existe evidencia de al menos dos péptidos que distingan inequívocamente entre dos eventos de splicing. Muchos estudios dan por identificadas isoformas alternativas cuando solo se ha podido identificar parte de ellas (Menon *et al.*, 2009; Menon & Omenn, 2010; Ly *et al.*, 2014). Como publicamos recientemente (Ezkurdia *et al.*, 2015a), en otros casos, aunque el método de identificación de la isoforma alternativa es correcto, el número de falsos positivos está subestimado (Kim *et al.*, 2014; Wilhelm *et al.*, 2014).

Durante el desarrollo de esta tesis he realizado diferentes contribuciones al campo de la proteogenómica. Así, he contribuido a la verificación de bases de datos genómicas a partir de experimentos proteómicos (Consortium & Consortium, 2012; Harrow *et al.*, 2012), a la búsqueda y validación de proteínas quiméricas (Frenkel-Morgenstern *et al.*, 2012), a la caracterización de splicing alternativo en humano y ratón (Ezkurdia *et al.*, 2012b; Abascal *et al.*, 2015a), al estudio de las isoformas principales (Ezkurdia *et al.*, 2015b), a establecer y representar el número de genes codificantes de proteína en el genoma humano (Ezkurdia *et al.*, 2014a) y a facilitar la utilización y acceso de resultados a la comunidad científica (Rodriguez *et al.*, 2012).

Entre las contribuciones mencionadas, para esta memoria he decidido incluir únicamente el trabajo relacionado con la proteogenómica del splicing aternativo (Ezkurdia *et al.*, 2012a; Abascal *et al.*, 2015a; Ezkurdia *et al.*, 2015b) y la búsqueda de proteínas quiméricas (Frenkel-Morgenstern *et al.*, 2012), así como las aportaciones a la definición de splicing en la enciclopedia de biofísica (Ezkurdia *et al.*, 2013) y a una revisión de proteómica dirigida

(Maiolica *et al.*). De esta forma, expongo los análisis proteogenómicos y de caracterización de splicing alternativo llevados a cabo utilizando dos grandes conjuntos de datos proteómicos a gran escala. Por último, muestro el análisis de representación de isoformas principales en humano.

2. Hipótesis y objetivos

El objetivo principal de esta tesis es estudiar el *splicing* alternativo a nivel de proteína, a partir del análisis de experimentos de proteómica a gran escala. Para ello se han definido los siguientes **objetivos específicos**:

1. Desarrollar estrategias computacionales que permitan el análisis de datos proteómicos a gran escala para la detección de isoformas alternativas

2. Realizar un estudio exploratorio de la importancia del fenómeno del *splicing* alternativo a nivel de proteínas en humano.

3. Evaluar la calidad de los resultados del estudio piloto, mediante la inspección manual del comportamiento estadístico de dichas estrategias en distintos conjuntos de proteómica.

4. Realizar un estudio a mayor escala de las isoformas generadas por *splicing* alternativo, en base a las estrategias desarrolladas en el estudio piloto.

5. Caracterizar funcional y estructuralmente los casos de variantes alternativas detectadas a nivel de proteína.

6. Realizar un análisis comparativo de las isoformas detectadas en experimentos de proteómica para distintos organismos modelo.

7. Investigar la propensión de los genes codificantes a presentar una o más isoformas dominantes.

3. Resultados

En esta sección se presentarán los análisis llevados a cabo a partir de los flujos de trabajo desarrollados para la detección de péptidos en experimentos de proteómica a gran escala (ver Materiales y Métodos).

La sección de resultados se ha dividido en tres subsecciones principales. En la primera mostraremos los resultados del análisis proteogenómico desarrollado a partir de dos versiones de la base de datos de GENCODE. En la segunda subsección presentaremos el análisis de caracterización de *splicing* alternativo a nivel de proteína. En esta sección se intercalan los resultados obtenidos en dos trabajos sucesivos para los que se utilizaron dos grandes conjuntos de datos proteómicos a gran escala: *CNIO_proteodata* y *Eight_proteodata* (ver Materiales y Métodos). La tercera subsección está dedicada al análisis de las isoformas principales a nivel de proteína.

3.1. Análisis proteogenómico

En el año 2003 un consorcio internacional comenzó con la identificación de todos los elementos funcionales del genoma humano en un proyecto llamado Encyclopedia of DNA Elements (ENCODE Project Consortium, 2011). Este proyecto en un principio se limitó a caracterizar un 1% del genoma (*Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*, 2007), ampliándose a todo el genoma en 2007. Como parte de esta iniciativa, GENCODE (Harrow *et al.*, 2006) comenzó con la anotación de las características del genoma humano centrándose inicialmente en el 1%. El objetivo del consorcio GENCODE es la identificación y caracterización completa del genoma humano (incluyendo, transcritos codificantes y no codificantes y pseudogenes) a partir de la combinación de análisis computacionales, anotaciones manuales y validación experimental. Como punto de partida GENCODE utilizó las predicciones automáticas de Ensembl (Hubbard *et al.*, 2002) para los cromosomas 21 y 22 para su revisión manual, progresivamente se fueron anotando el resto de cromosomas. Desde la publicación de la primera versión de estos datos, se han añadido algunas regiones codificantes nuevas y un

número cada vez mayor de nuevos transcritos de *splicing* alternativo. Los datos de GENCODE están accesibles en gencodegenes.org y a través de los navegadores de ENSEMBL y UCSC.

Como nuestro grupo formaba parte del consorcio de GENCODE nos planteamos hasta que punto se podían utilizar datos proteómicos para validar las anotaciones genómicas.

3.1.1. Validación de anotaciones genómicas

Para la validación de anotaciones genómicas se utilizaron los resultados generados a partir de las búsquedas de los péptidos de GENCODE en experimentos de proteómica. Este conjunto de experimentos fue extraído de dos repositorios públicos de proteómica y nos vamos a referir a el como *CNIO_proteodata* (ver Materiales y Métodos).

Las anotaciones manuales de los transcritos de GENCODE se dividían en tres categorías principales basadas en la evidencia relativa a su anotación: *Known*, formada por secuencias conocidas que se pudieron mapear a proteínas que proceden de Swiss-Prot; *novel*, CDS que compartían con un CDS conocido al menos un 60% de longitud de secuencia o la misma estructura de dominios; y, por último, los anotados como *putative*, que eran aquellos que compartían menos del 60% de longitud con un CDS conocido (en la versión actual de GENCODE no existe esta categoría). El análisis se hizo con el objetivo de validar los genes y las secuencias codificantes (CDS) de GENCODE que estaban definidos como *novel* y *putative*.

Tenemos que tener en cuenta que a pesar de que las anotaciones de GENCODE abarcan todo el genoma humano, éstas no son definitivas. Esto implica que no vamos a poder detectar aquellos genes codificantes y variantes alternativas que no hayan sido anotadas, o los polimorfismos de un solo nucleótido.

Estableciendo un umbral del 1% de FDR, identificamos 75.474 péptidos trípticos que mapeaban a 7.597 genes de los 22.304 anotados en la versión 3C de GENCODE (el 34% de

los genes). Entre los genes identificados, encontramos péptidos que mapeaban inequívocamente a siete genes con transcritos que aparecían anotados como *putative* o *novel*. Se pudieron validar 12 transcritos, cinco que estaban anotados como *putative,* y nueve que estaban anotados como *novel*. En las **Figuras 2** y **3** se ilustran dos ejemplos.

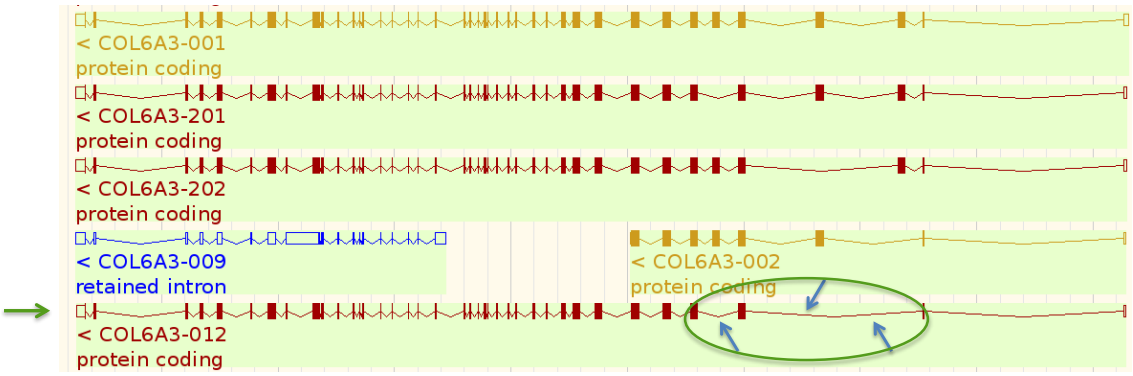


Figura 2. Ejemplo ilustrativo de isoforma anotada como *novel*.

Utilizando el navegador del genoma de Ensembl, podemos ver un parte del modelo del gen *COL6A3*. El transcrito ENST00000472056 (COL6A3-012), que aparece anotado como *novel*, esta indicado con una flecha verde. El círculo verde, indica la región que difiere en el transcrito ENST00000472056, en la zona del 3' terminal. Las posiciones de los tres exones que no están no presentes en este transcrito, pero que aparecen anotados en otros transcritos están señalados con flechas azules.

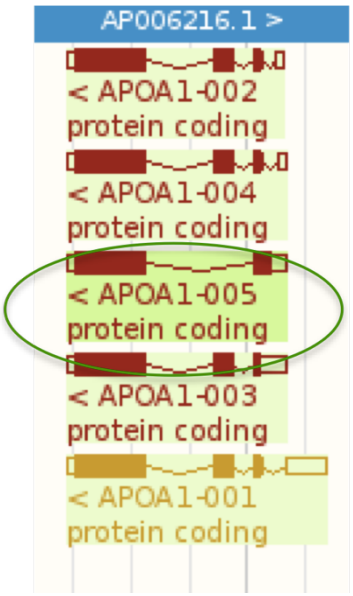


Figura 3. Ejemplo ilustrativo de isoforma anotada como *putative*.

Utilizando el navegador del genoma de Ensembl, podemos ver la anotación para el gen *APOA1*. El círculo verde marca el transcrito ENST00000375329, que está anotado como *putative* para este gen. En este caso, el transcrito codifica para un N-terminal diferente al resto de las isoformas.

La última versión de GENCODE7 se publicó mientras trabajábamos en este estudio y decidimos repetir los análisis utilizándola. En el caso de GENCODE7, fuimos capaces de identificar el 39,1% de los genes anotados mapeando 83.054 péptidos trípticos a 8.098 genes de los 20.681 que aparecen anotados en la base de datos. En este caso detectamos 33 transcritos anotados como *putative* y 50 anotados como *novel*.

También se analizaron los genes anotados como *pseudogene*. En este caso, se detectaron péptidos que mapeaban a 15 genes que aparecían anotados como pseudogenes. Algunos de estos genes, y como consecuencia de este trabajo, han sido reclasificados posteriormente como genes codificantes, como en el caso de *ZNF66P*, *IGLC6* y *MSTP9*.

3.1.2. Evidencias de traducción para ARN mensajeros con degradación mediada por mutaciones terminadoras

La degradación del ARN mensajero mediada por mutaciones terminadoras o NMD (Nonsense mediated decay), es un mecanismo celular encargado de degradar transcritos mRNA anormales que contienen codones de terminación prematuros (BehmAnsmant *et al.*, 2007; Isken & Maquat, 2008; Chang *et al.*, 2010) y que se estima afecta al 75-90% de los genes humanos (McGlincy & Smith, 2008). El proceso de activación de NMD solo se entiende parcialmente y algunos trabajos recientes sugieren que podría haber diferencias sustanciales entre los mecanismos de NMD para diferentes organismos (BehmAnsmant *et al.*, 2007; Brogna & Wen, 2009). Es interesante que la existencia de expresión para transcritos candidatos a NMD (Barberan-Soler *et al.*, 2009; Filichkin *et al.*, 2009) ha sido demostrada.

De los 72.731 transcritos que había en GENCODE 3C 3.939 estaban anotados como dianas potenciales de NMD. Estas asignaciones se hacen automáticamente para todos los transcritos que finalizan antes de 50 nucleótidos a partir de un punto de *splicing*. En este análisis se detectaron por primera vez cuatro isoformas que estaban etiquetadas para degradación por NMD.

En la **Figura 4**. Podemos ver un ejemplo en la detección del péptido EYNTGNSNDPGQR que mapea el transcrito ENST00000425608 del gen *RBM6*. Este transcrito se salta un exón dando lugar a un cambio del marco de lectura y a un codón de parada prematuro. El péptido EYNTGNSNDPGQR mapea a la región del transcrito donde se unen los exones anterior y posterior al exón perdido. Se encontraron un total de cinco asignaciones espectro-péptido (PSM, del inglés Peptide-Spectrum Match) que validaron la existencia del péptido en experimentos diferentes.

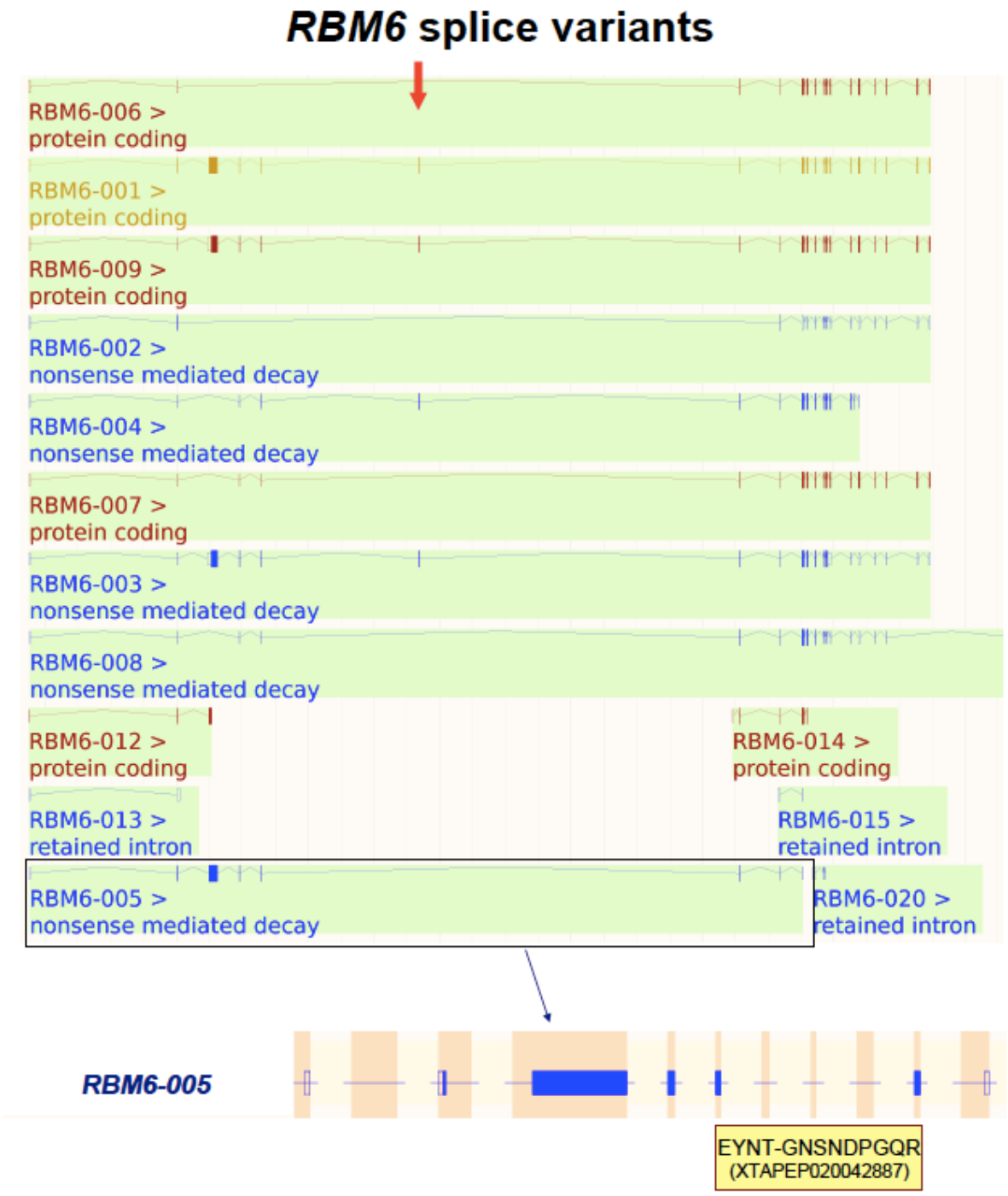


Figura 4. Anotación del gen *RBM6*.

La anotación para el gen *RBM6* en el navegador de Ensembl mostrando el péptido para la isoforma RBM6-005 (ENST00000425068). Este gen consta de 22 variantes distintas. El transcrito RBM6-005 que se salta el exón que señala la flecha, dando lugar a un cambio de marco, y un codón de parada prematuro. El péptido encontrado (mostrado abajo) está

separado por un guion (-) que marca la zona de unión del exón constitutivo y el exón que cambia de marco.

3.1.3. Proteínas quiméricas

Los ARNm quiméricos se producen a partir de la unión de exones de dos o más genes (Pirrotta, 2002; Horiuchi & Aigaki, 2006). La traducción de estos ARNm podría dar lugar a una proteína nueva. Se ha identificado un número elevado de transcritos quiméricos a partir de datos de ESTs y de experimentos de secuenciación masiva de ARN. Las proteínas quiméricas se generan a partir de procesos complejos como la translocación cromosómica, la duplicación en tándem o la retrotransposición.

Existen pocos casos en los que se haya podido demostrar la traducción de estos ARNm en proteínas. En su mayoría, estos casos son el resultado de translocaciones cromosómicas y están asociados con cáncer. El ejemplo de la proteína de fusión bcr-abl es uno de los casos más estudiados y está asociado a la leucemia crónica mieloide (Raitano *et al.*, 1995).
Para detectar si existían quimeras expresadas a nivel de proteína, fue necesario generar una nueva base de datos de búsqueda. Se generaron 32.382 proteínas a partir de la traducción de los seis posibles sentidos de los marcos abiertos de lectura de 5.397 ARNm (ver Materiales y Métodos).

Los péptidos detectados con un 1% de FDR (utilizando como base de datos las quimeras sobre *CNIO_proteodata*) se filtraron en base a los siguientes criterios:

- Se descartaron todos los péptidos que pertenecen a proteínas anotadas en GENCODE 3C.
- Los péptidos tenían que corresponder con la zona de unión de dos genes diferentes y contener al menos tres aminoácidos pertenecientes a cada gen.
- Se filtraron todos aquellos péptidos que pudieran explicarse a partir de una mutación puntual (SNP, del inglés Single Nucleotid Polimorfism) de un péptido anotado.

Tan sólo se pudieron detectar dos péptidos que cumplían los requisitos de validación. El péptido LWTVSRCLTASHTVPIYEGYALPHAILR y su subpéptido ASHTVPIYEGYALPHAILR. El péptido contiene 12 aminoácidos que solapan con la zona de unión de los dos genes del ESTid "BM838228.1", la proteína ribosomal *RPL13A* y la actina *ACTG1*. Este péptido se detectó en 12 experimentos distintos de PeptideAtlas.

Dado que se trataba de un caso único, se decidió corroborar el resultado utilizando más experimentos proteómicos. El grupo de Levin del instituto Weizmann llevó a cabo las mismas búsquedas utilizando como muestras tres líneas celulares humanas (MCF7, OVCAR-3 y DU-145) y usando el espectrómetro de masas en modo de adquisición independiente de datos (Levin *et al.*, 2011). Las búsquedas llevadas con ProteinLynx Global Server (IdentityE) detectaron 14 péptidos correspondientes a otras 11 proteínas quiméricas. Los 14 péptidos detectados junto al péptido ASHTVPIYEGYALPHAILR encontrado a partir de nuestro flujo de trabajo han sido sintetizados y analizados por el mismo grupo utilizando proteómica dirigida SRM. Como resultado se pudo confirmar la presencia del péptido detectado a partir de nuestro flujo de trabajo ASHTVPIYEGYALPHAILR y los péptidos GRLGQPAMAK y VISSIEQKTMAAPSVK que fueron detectados por el grupo de Levin.

3.2. Detección y caracterización de isoformas de *splicing* alternativo

En esta sección se presentarán los análisis para la caracterización de *splicing* alternativo correspondientes a dos trabajos sucesivos realizados sobre los conjuntos de datos proteómicos de *CNIO_proteodata* y *Eight_proteodata* respectivamente, utilizando las bases de datos de GENCODE 3C y GENCODE 20. El conjunto de datos proteómicos *CNIO_proteodata* se generó a partir de los experimentos proteómicos de dos repositorios públicos y el conjunto *Eight_proteodata* que incluye también el conjunto *CNIO_proteodata*, a partir de dos repositorios públicos y de los resultados obtenidos en seis trabajos (ver Materiales y Métodos).

Cuando se desarrollaron los análisis para el primer trabajo, la mayoría de las publicaciones de datos de proteómica utilizaban bases de datos que no contenían isoformas alternativas o contenían un número pequeño de las isoformas anotadas. Para poder realizar los análisis sobre el conjunto de datos *CNIO_proteodata* fue necesario repetir las búsquedas utilizando bases de datos más actualizadas en colaboración con GENCODE. En el momento de la realización del estudio ampliado, la mayoría de datos proteómicos publicados empezaban utilizar bases de datos con anotaciones completas de isoformas. Esto nos permitió ampliar la cantidad de datos experimentales considerablemente, sin tener que repetir las búsquedas. De hecho, dada la magnitud del conjunto de datos *Eight_proteodata* repetir las búsquedas contra otra base de datos hubiera excedido nuestra capacidad computacional (ver Materiales y Métodos).

3.2.1. Isoformas alternativas en humano

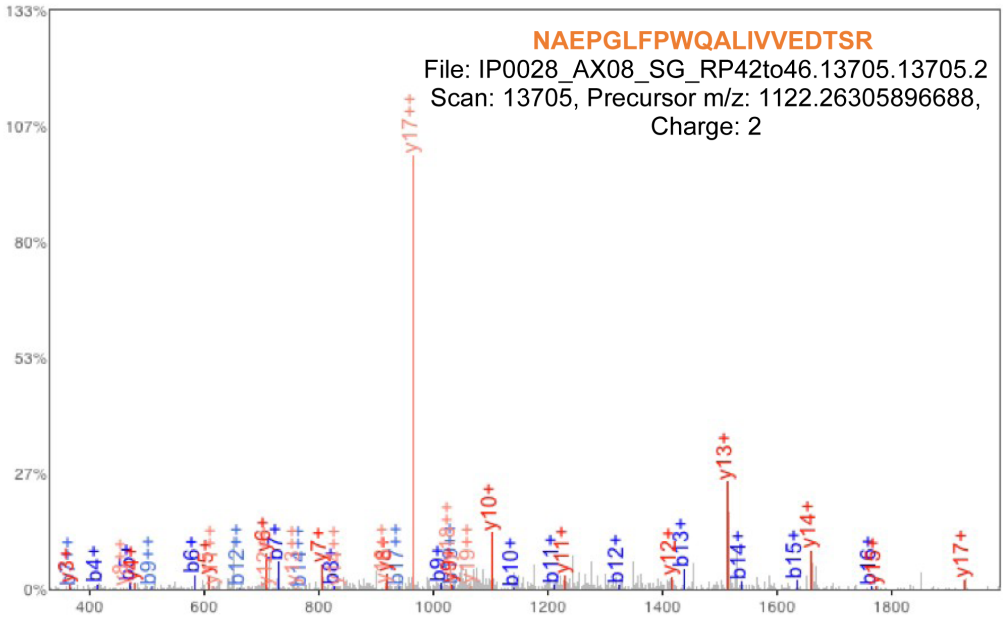
Para la detección de isoformas alternativas a nivel de proteína, se tuvieron en cuenta únicamente aquellos péptidos que distinguían inequívocamente al menos dos transcritos diferentes expresados por un mismo gen. En la Figura 5, se ilustra un ejemplo del mapeo de péptidos realizado. A partir del mapeo del péptido NAEPGLFPWQALIVVEDTSR, podemos distinguir inequívocamente las isoformas ENST00000296280 y ENST00000392472 (la secuencia de esta isoforma no se incluye en la figura); y con el

péptido FPETLMEIEIPIVDHSTCQK, la isoforma ENST000000337774. Aunque no podemos saber cual de las dos primeras se está expresando, sabemos que el gen *MASP1* expresa al menos dos isoformas. En la Figura 6 podemos ver otro ejemplo, en el que además hemos podido mapear los péptidos en estructuras homólogas.

En el caso de las deleciones, es necesario encontrar un péptido que abarque la intersección entre los exones adyacentes al eliminado. Si se trata de una sustitución o una inserción, basta con que el péptido mapee a alguna de sus intersecciones.

A >ENST00000296280|ENSG00000127241|MASP1-003|3|protein_coding

MRWLLLYYALCFSLSKASAHTVELNNMFGQIQSPGYPDSYPSDSEVTWNITVPDGFRIKLYFMHFNLESSYLCEYDYVKVETEDQV
LATFCGRETTDTEQTPGQEVVLSPGSFMSITFRSDFSNEERFTGFDAHYMAVDVDECKEREDEELSCDHYCHNYIGGYYCSCRFGY
ILHTDNRTCRVECSDNLFTQRTGVITSPDFPNPYPKSSECLYTIELEEGFMVNLQFEDIFDIEDHPEVPCPYDYIKIKVGPKVLGP
FCGEKAPEPISTQSHSVLILFHSDNSGENRGWRLSYRAAGNECPELQPPVHGKIEPSQAKYFFKDQVLVSCDTGYKVLKDNVEMDT
FQIECLKDGTWSNKIPTCKIVDCRAPGELEHGLITFSTRNNLTTYKSEIKYSCQEPYYKMLNNNTGIYTCSAQGVWMNKVLGRSLP
TCLPECGQPSRSLPSLVKRIIGGRNAEPGLFPWQALIVVEDTSRVPNDKWFGSGALLSASWILTAAHVLRSQRRDTTVIPVSKEHV
TVYLGLHDVRDKSGAVNSSAARVVLHPDFNIQNYNHDIALVQLQEPVPLGPHVMPVCLPRLEPEGPAPHMLGLVAGWGISNPNVTV
DEIISSGTRTLSDVLQYVKLPVVPHAECKTSYESRSGNYSVTENMFCAGYYEGGKDTCLGDSGGAFVIFDDLSQRWVVQGLVSWGG
PEECGSKQVYGVYTKVSNYVDWVWEQMGLPQSVVEPQVER



B >ENST00000337774|ENSG00000127241|MASP1-001|3|protein_coding

MRWLLLYYALCFSLSKASAHTVELNNMFGQIQSPGYPDSYPSDSEVTWNITVPDGFRIKLYFMHFNLESSYLCEYDYVKVETEDQV
LATFCGRETTDTEQTPGQEVVLSPGSFMSITFRSDFSNEERFTGFDAHYMAVDVDECKEREDEELSCDHYCHNYIGGYYCSCRFGY
ILHTDNRTCRVECSDNLFTQRTGVITSPDFPNPYPKSSECLYTIELEEGFMVNLQFEDIFDIEDHPEVPCPYDYIKIKVGPKVLGP
FCGEKAPEPISTQSHSVLILFHSDNSGENRGWRLSYRAAGNECPELQPPVHGKIEPSQAKYFFKDQVLVSCDTGYKVLKDNVEMDT
FQIECLKDGTWSNKIPTCKIVDCRAPGELEHGLITFSTRNNLTTYKSEIKYSCQEPYYKMLNNNTGIYTCSAQGVWMNKVLGRSLP
TCLPVCGLPKFSRKLMARIFNGRPAQKGTTPWIAMLSHLNGQPFCGGSLLGSSWIVTAAHCLHQSLDPEDPTLRDSDLLLSPSDFKI
ILGKHWRLRSDENEQHLGVKHTTLHPQYDPNTFENDVALVELLESPVLNAFVMPICLPEGPQQEGAMVIVSGWGKQFLQRFPETLM
EIEIPIVDHSTCQKAYAPLKKKVTRDMICAGEKEGGKDACAGDSGGPMVTLNRERGQWYLVGTVSWGDDCGKKDRYGVYSYIHHNK
DWIQRVTGVRN

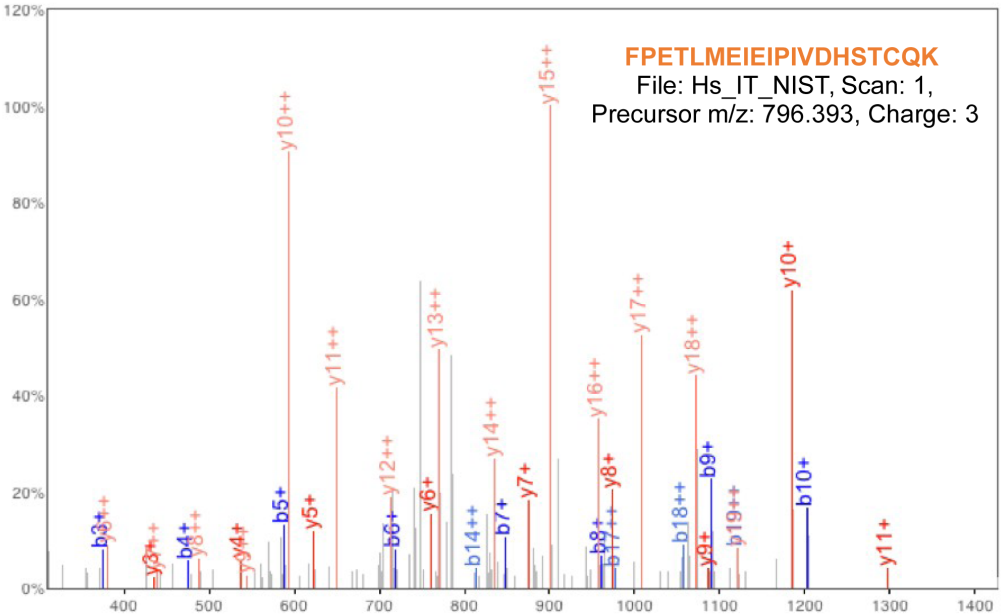


Figura 5. Mapeo de péptidos en *MASP1*

a) Secuencia de las isoformas MAP1-003 con el péptido detectado que discrimina a esta isoforma ya MAP1-005 (no mostrado) del resto de isoformas en naranja. b) Espectro del péptido NAEPGLFPWQALIVVEDTSR c) Secuencia de la isoforma MAP1-001 con el péptido que discrimina esta isoforma (FPETLMEIEIPIVDHSTCQK) en naranja y abajo el espectro del péptido.

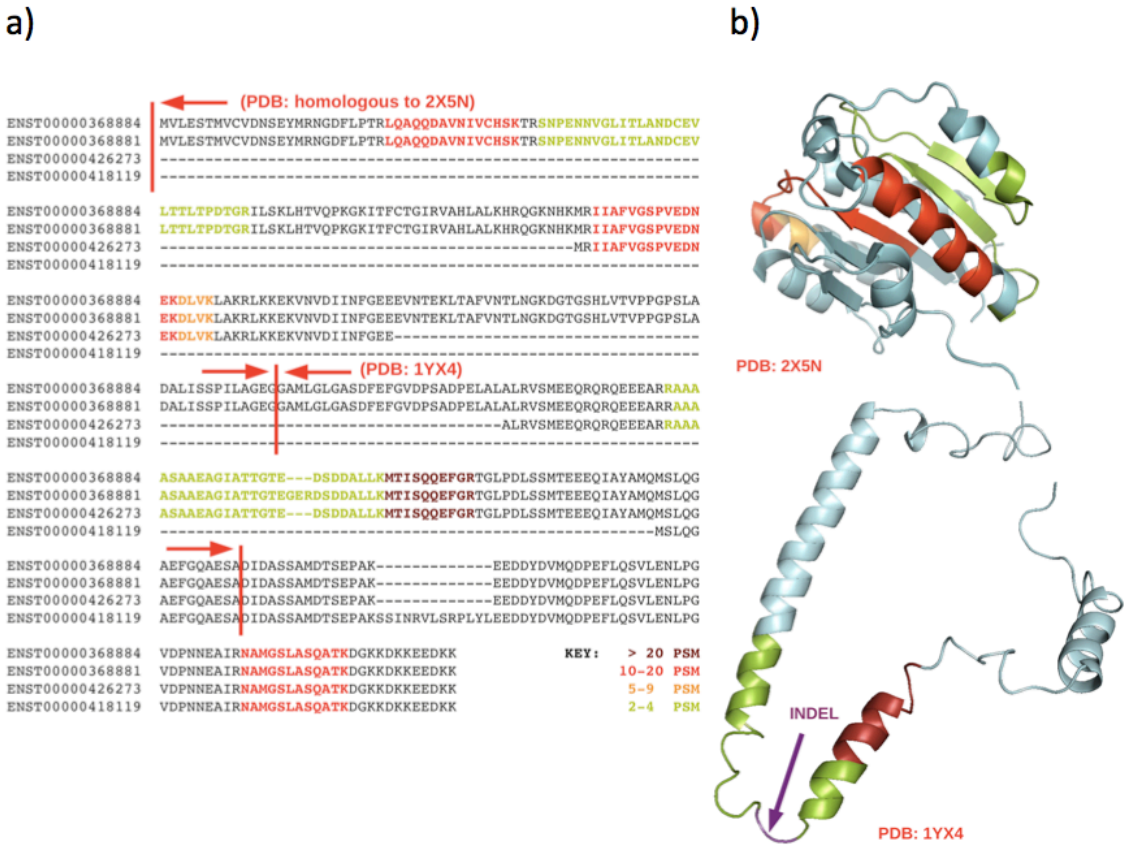


Figura 6. Ejemplo de mapeado de péptidos. **A)** Mapeo de péptidos para las cuatro isoformas alternativas anotadas del gen *PSMD4*. Los péptidos detectados en los experimentos de proteómica aparecen coloreados en función del número de PSM encontrados. Para identificar un evento de *splicing* era necesario encontrar péptidos para los dos lados de un evento. Por lo tanto a partir de estos péptidos se pueden distinguir las isoformas ENST00000368884 y ENST00000368881 que difieren en un indel de 3 residuos. **B)** Los péptidos detectados se han mapeado a la estructura homóloga del PDB (Berman *et al.*, 2000) 2X5N, con una homología del 50% con respecto al N-terminal de *PSMD4* y a la estructura 1YX4 que abarca la zona

central de las isoformas. La flecha indica la zona de la estructura 1YX4 donde se insertarían los tres residuos de la isoforma ENST00000368881. Esta zona, al igual que gran parte de la estructura 1YX4, corresponde a una región desordenada.

En el análisis de isoformas alternativas sobre el conjunto *CNIO_proteodata* (utilizando la base de datos GENCODE 3C) se detectaron 150 genes que expresaban dos o más isoformas. Entre estos genes, había 13 que expresan tres o más proteínas.

Algunos de los genes para los que se encontraron dos o más isoformas alternativas, habían sido ampliamente caracterizados en la literatura, como es el caso de las isoformas p16INK4a y p14ARF del gen *CDKN2A* (NormanSharpless, 2005). Entre los genes que se detectaron expresando más de una isoforma, tenemos un total de cinco componentes principales del nucleosoma, Lamin A/C, LAP2, Proteína 4.1R, titin y la espectrina alfa-II (Simon & Wilson, 2011). Para Lamin A/C, el mayor componente de laminas nucleares junto con la timopoyetina (*TMPO*), se detectaron dos isoformas conocidas, LAMIN A y LAMIN C. También se encontró evidencia para dos isoformas de timopoyetina, LAP2-alpha y LAP2-beta y dos isoformas de otra proteína que participa en la unión del citoesqueleto, la proteína 4.1R (*EPB41*), que ha sido relacionada con lamin A/C y LAP2-alpha (Meyer *et al.*, 2011).

Además se pudo detectar evidencia de expresión de distintas isoformas alternativas para los genes *CUX1*, *NEBL* y *MACF1*. Estos tres genes generan isoformas que comparten parte de su secuencia proteica pero divergen considerablemente en la parte restante, incluyendo dominios funcionales diferentes bien caracterizados. El gen *CUX1* genera dos isoformas, CDP y CASP con C-terminales diferentes (Lievens *et al.*, 1997). El C-terminal de CDP contiene tres copias del dominio de unión a ADN CUT. En el caso de la isoforma CASP, cuya función está relacionada con transporte vesicular (Gillingham *et al.*, 2002), estos dominios son reemplazados por un solo dominio CASP_C. El gen *MACF1* (Microtubule-actin cross-linking factor 1) genera dos isoformas que difieren en el N-terminal. La región N-terminal de la isoforma 1 tiene dos dominios CH mientras que la isoforma 4 contiene varios dominios de plectina. El gen *NEBL* (nebulette) también difiere en el N-terminal y tiene

una isoforma con varias repeticiones de *nebulin* y otro donde son reemplazados por el dominio *LIMS*.

En el estudio ampliado se utilizaron los péptidos provenientes de ocho experimentos proteómicos a gran escala, con el objetivo de abarcar el máximo número posible de tejidos y líneas celulares. Para lograr unos resultados fiables se aplicaron al conjunto de datos proteómicos *Eight_proteodata* una serie de filtros de validación como exigir que fueran péptidos trípticos o que hubieran sido detectados por más de un motor de búsqueda (ver Materiales y Métodos). Con este método se consiguió identificar el 63,9% (12.716) de los genes humanos inequívocamente. De los 149.954 péptidos obtenidos, 111.382 fueron capaces de discriminar entre las diferentes isoformas en la base de datos GENCODE 20. Se encontraron 246 genes que expresan más de una isoforma, un 60% más que en el estudio anterior.

En la comparación entre ambos estudios, detectamos 77 genes en común entre ellos. Cuando analizamos las discrepancias entre los dos estudios observamos que estas se debían principalmente a la forma de validar los péptidos y a las diferencias entre las versiones de GENCODE utilizadas. En el segundo estudio los requisitos de validación fueron más estrictos, en parte porque se incluyeron más experimentos y ello implicaba un aumento de falsos positivos (ver Materiales y Métodos). Algunos de los genes detectados en el estudio piloto también se encontraron en el estudio ampliado pero no pasaron el umbral del filtro aplicado. En otras ocasiones, estas discrepancias se deben a problemas relacionados con la anotación del gen. En muchos casos, el gen para el que se habían encontrado dos isoformas resultó ser en realidad dos genes diferentes cuya anotación estaba actualizada en la versión de GENCODE usada en el segundo estudio. Por ejemplo, los genes *TAF9, PRSS1 y MT1X* que daban lugar a más de una isoforma cuando utilizamos GENCODE 3C como referencia, pasaron a estar anotados como dos genes en GENCODE 20.

3.2.2. Detección de isoformas alternativas en ratón y mosca del vinagre

Con el fin de establecer las comparaciones con los datos obtenidos en humanos también se analizaron los proteomas de ratón y mosca.

En el primer estudio, el análisis de isoformas alternativas para ratón se hizo a partir de las anotaciones de la versión NCBIM37.61 de Ensembl, utilizando datos de PeptideAtlas (ver Materiales y Métodos). Fueron detectadas isoformas de *splicing* alternativo en 49 genes. Para el análisis en ratón del estudio ampliado utilizamos la base de datos de NIST, PeptideAtlas y la base de datos generada para el primer estudio (ver Materiales y Métodos). Se detectaron 56 genes que daban lugar a 68 eventos de *splicing*.

La detección de un número menor de isoformas es debida al tamaño de la base de datos de anotaciones genómicas y al número de experimentos de proteómica analizados. Hay que tener en cuenta que las bases de datos de anotaciones genómicas para ratón no son tan completas, por lo que el número de isoformas alternativas detectables es menor. Por otro lado, la cantidad de experimentos proteómicos sobre las que se realizaron la búsquedas es mucho menor.

En el caso de la mosca del vinagre, se utilizaron 130 genes para los que se detectaron isoformas alternativas expresadas a nivel de proteína en el trabajo de Tress et al. (Tress *et al.*, 2008). Estas detecciones se llevaron a cabo a partir de dos experimentos de proteómica a gran escala en los que se analizaron un total de 8.166 genes.

3.2.3. Factores que influyen en la detección de isoformas

En este punto se muestran los factores que inciden en la detección de genes que expresan más de una *isoforma*. Hay tres factores que se han tenido en cuenta a la hora de evaluar hasta qué punto son detectables las isoformas alternativas: los niveles de expresión, las características de las secuencias y la longitud de las mismas.

Las diferencias a nivel de secuencia entre las isoformas anotadas y sus longitudes, son factores que influyen en su detectabilidad. La probabilidad de mapear más de una isoforma por gen está relacionada con la longitud de secuencia de sus variantes de *splicing* y la identidad de secuencia entre ellas. La detección de una isoforma requiere al menos el mapeo inequívoco de un péptido. Evidentemente, será más fácil detectar dos isoformas completamente diferentes entre sí, que dos isoformas cuya única diferencia estriba en la inserción de un aminoácido.

Para medir la predisposición teórica de detección de isoformas alternativas en el primer estudio, se tuvo en cuenta la longitud de la proteína, el número de variantes de *splicing* anotadas y el número de exones. Esta simulación *in silico* se realizó a partir de tres conjuntos de genes. El primero comprende los genes detectados que expresan isoformas alternativas: *Genes IA*. El segundo está formado por genes que tienen anotadas varias isoformas pero sólo se detectó una: *Genes Background*. El tercero está formado por aquellos genes que tienen anotadas varias isoformas pero no se detectó ninguna: *Genes ND*.

El calculo de la probabilidad de detección de una isoforma, se realizó a partir de una aproximación de simulación aleatoria. Primero se realizó la digestión *in silico* de los péptidos de todos los genes anotados en GENCODE. Se eligieron 14,000 péptidos al azar (este número se calibró para poder identificar un número de genes similar al del análisis, 8,000 genes) y se calculo el número de genes que se habrían identificado con al menos dos isoformas alternativas. El experimento se repitió 1,000 veces y se asignó a cada gen una probabilidad teórica de detección en base al porcentaje de veces en las que al menos dos de sus isoformas eran identificadas.

Para calcular la influencia en la detección de los niveles de expresión, se utilizó la base de datos HuGE (Haverty *et al.*, 2002) como referencia. Se extrajeron los niveles de expresión de 2.512 genes del conjunto de *Genes Background* y de 109 genes del conjunto *Genes IA* para los que había información. Para cada uno de los genes se calculó la media aritmética en cada tejido y la media global de expresión a partir de éstas. La expresión a nivel de proteína

de los experimentos, se estimó contando el número de PSM inequívocos para cada uno de los genes.

	Longitud media del CDS	Variantes totales	Exones totales	Expresión (HUGE)	PSM	Probabilidad de detección
Genes ND	450,9	4,19	11,8	77,21	-	15,89
Genes Background	488,82	4,52	14,64	190,82	106,96	18,2
Genes IA	602,38	7,63	20	323,71	754,66	32,57

Tabla 2. Características de los conjuntos de genes anotados con más de una isoforma en función de su probabilidad de detección.

Los resultados se muestran en la Tabla 2. Los genes del grupo *Genes Background* son un poco más largos que los del grupo *Genes ND* además de tener más variantes de *splicing* y de exones por gen. Podemos observar además una gran diferencia en los niveles medios de expresión de la base de datos HuGE, donde la media de expresión del conjunto *Genes Background* es más del doble que la de los 7,329 genes del conjunto de *Genes ND* (190.8 a 77.2). Como cabría esperar, las proteínas detectadas en los experimentos de proteómica están más expresadas que aquellas que no han sido detectadas.

Las diferencias entre el conjunto de *Genes Background* y el conjunto de *Genes IA* son aún más grandes. En este caso, podemos observar que la longitud de las proteínas del conjunto de *Genes IA* tienen de media 110 residuos más. El número de exones y de variantes anotadas por gen también aumenta, así como los niveles de expresión a nivel de transcrito. Asimismo, aumenta la probabilidad teórica media de detección que pasa de un 33% para el conjunto *Genes IA* a un 18,2% para el conjunto *Genes Background*. El número medio de péptidos por gen encontrados es siete veces mayor que el encontrado en el conjunto *Genes Background*.

Observando los datos obtenidos en la Tabla 2, pudimos concluir que nuestros análisis detectaban aquellas variantes de *splicing* más fáciles de detectar.

3.2.4. Caracterización de splicing para el primer estudio

3.2.4.1 Exones homólogos

Para estudiar este tipo de *splicing*, definimos como exones homólogos equivalentes los exones homólogos al exón que reemplazan. Muchas de estas isoformas homólogas equivalentes (HE) están generadas a partir de exones mutuamente excluyentes (MEE). La definición de MEE es más estricta porque implica que los dos exones mutuamente excluyentes no pueden aparecer en el mismo transcrito y además tienen que ser internos, no pueden estar en los extremos 3' o 5' del transcrito. En las HE no siempre se cumple esta regla, a veces los dos exones pueden estar en el mismo transcrito como en el caso de *TPM1*. MEE es una forma de *splicing* poco frecuente (Tress *et al.*, 2007c; Wang *et al.*, 2008).

Un ejemplo de isoformas generadas a partir de exones mutuamente excluyentes lo encontramos en dos isoformas producidas por el gen de ketohexokinasa, *KHK*, que expresa una enzima que cataliza la fosforilación de la fructosa. Estas isoformas difieren en un solo exón homólogo que se intercambia. Para este ejemplo concreto, las estructuras de ambas isoformas están resueltas en el PDB y nos permiten analizar cómo afectaría este evento a su estructura. La región de la estructura de la proteína cubierta por este sitio de unión incluye los residuos de unión a sustrato, así como un residuo catalítico (Figura 7). A pesar de que muchos de los residuos de unión a fructosa difieren en las dos isoformas, ambas se unen a la fructosa (Trinh *et al.*, 2009) y las predicciones llevadas a cabo por el servidor firestar (Lopez *et al.*, 2011) sugieren que el residuo catalítico afectado seguiría siendo funcional.

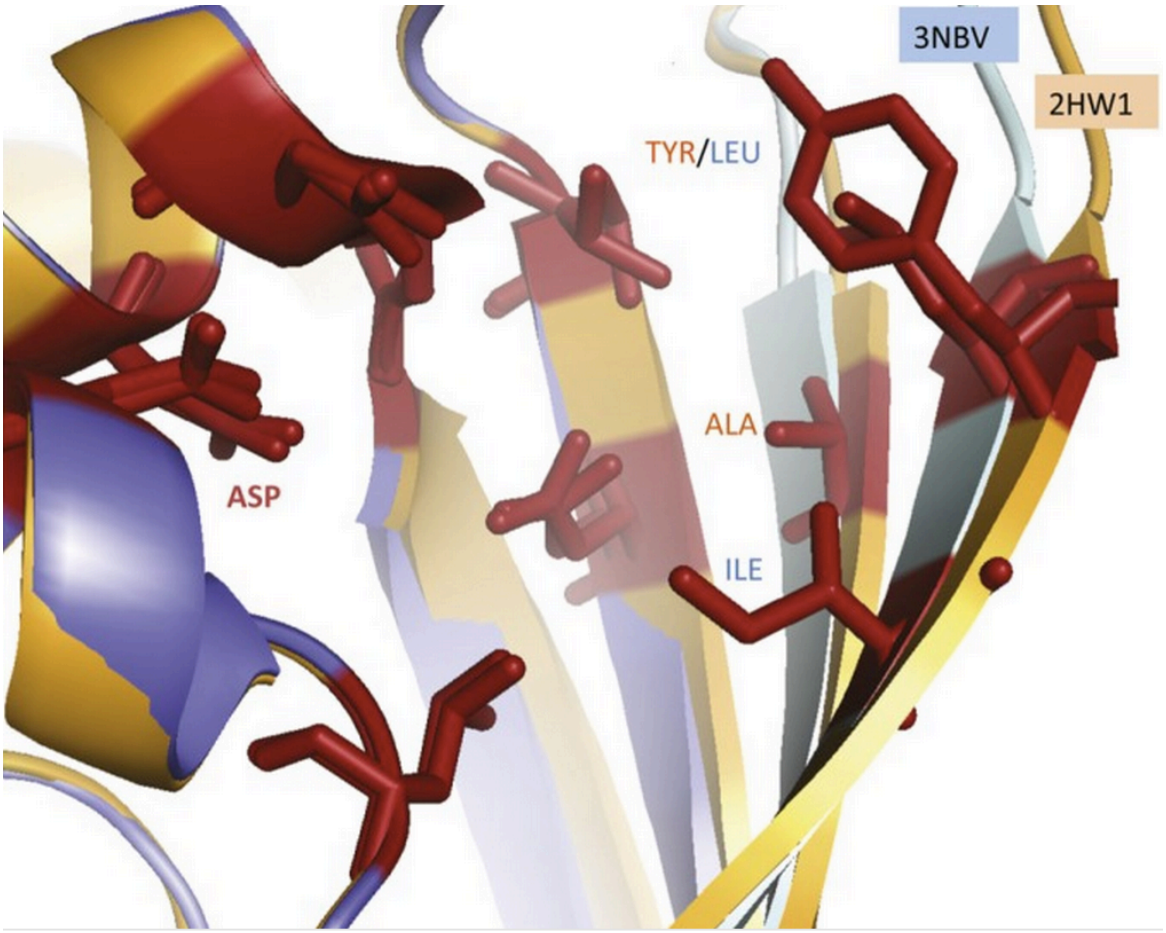


Figura 7. Alineamiento estructural del sitio activo para las dos isoformas alternativas detectadas para el gen ketohexokinasa, *KHK* en el estudio piloto (*CNIO_proteodata*). La estructura de la zona de unión periférica a fructosa de la isoforma cuya estructura esta resuelta con el código PDB 3NBV aparece alineada con la zona de unión central a fructosa de la isoforma cuya estructura esta resuelta con el código PDB 2HW1. Los residuos más cercanos al sustrato de fructosa para cada una de las isoformas aparecen dibujados como *sticks*. Se puede observar que los residuos que forman parte del sitio de unión tienen prácticamente la misma orientación con la excepción de los residuos pertenecientes al exón homólogo. Los residuos que están en contacto con la fructosa son diferentes en ambas isoformas, alanina y tirosina para 2HW1, y leucina e isoleucina para 3NBV. A pesar de estas diferencias, la hidrofobicidad de los residuos es similar y ambas isoformas fueron cristalizadas unidas a fructosa además de tener ambas los residuos catalíticos conservados (Lopez *et al.*, 2011).

En el primer estudio encontramos isoformas alternativas generadas a partir de exones homólogos en 19 genes, incluyendo *TPM1* para el que identificamos al menos cuatro isoformas diferentes de *splicing*. Este gen tiene nueve variantes anotadas en GENCODE que difieren entre sí en los exones 5' y 3', y en dos conjuntos de exones internos. El exón 3' y los exones internos son homólogos en secuencia y sufren *splicing* alternativo de manera mutuamente exclusiva. En total encontramos 572 PSM para los 28 péptidos inequívocos que mapean a las diferentes variantes de TPM1. Se detectaron PSMs para cada exón en las cuatro regiones de *splicing* alternativo, con lo que se pudo confirmar la presencia de cuatro eventos de *splicing*. Sin embargo, en teoría podrían estar traduciéndose las nueve isoformas que están anotadas para el gen. En el estudio ampliado se detectaron 60 genes HE y se realizaron varias pruebas de validación como mostramos más adelante.

3.2.4.2. *Splicing* NAGNAG

En el trabajo de Tanner (Tanner *et al.*, 2007b) se encontraron ocho pares de isoformas alternativas con deleciones de un solo aminoácido. En nuestro primer estudio, detectamos 11 genes que expresan isoformas alternativas que difieren solamente en la inserción o deleción de un único aminoácido. Se pudieron confirmar 10 pares de isoformas este tipo anotados en la base de datos TASSDB (Hiller *et al.*, 2007) que han sido generados por aceptores NAGNAG o donantes GYNGYN. Además de estos 11 genes, encontramos otros seis pares de isoformas alternativas generadas a partir de inserciones o deleciones de dos, tres y cuatro residuos; cinco de estas deleciones también aparecen confirmadas en la base de datos TassDB. También detectamos cinco genes con isoformas alternativas (*CUX1*, *IMMT*, *HNRNPR*, *HNRNPK* and *RBM26*) generadas por NAGNAG *splicing* y otros eventos de *splicing*.

3.2.4.3. Ribonucleoproteínas Heterogéneas-Nucleares

Dentro de los genes que aparecen anotados en el grupo de unión a ARN según el análisis de enriquecimeinto de términos GO realizado con la herramienta DAVID (Huang *et al.*, 2008) destacan 10 genes clasificados como Ribonucleoproteínas Heterogéneas-Nucleares

(hnRNPs). Estos genes están implicados en la regulación de *splicing* alternativo (Martinez-Contreras *et al.*, 2007; Venables *et al.*, 2008). De los 26 genes hnRNP anotados en la versión GENCODE 3C que expresan isoformas alternativas, detectamos péptidos para 23 de ellos. Asumiendo que la detección de isoformas alternativas para cualquier gen del conjunto *Genes Background* es equiprobable, esperaríamos haber encontrado isoformas para 0,62 genes hnRNP en el conjunto *Genes IA*. En el trabajo de Tanner (Tanner *et al.*, 2007b), observaron resultados similares (entre los 15 genes detectados que expresaban más de una isoforma, dos eran hnRNPs).

3.2.5. Significado estadístico de eventos de *splicing* en el primer estudio

Para poder cuantificar la expresión de las isoformas detectadas se creó un marcador a partir del recuento del número de PSM inequívocos. Este marcador, que llamaremos *AP Score*, se calcula a partir de los PSM de aquellos genes que tienen evidencia de expresar dos isoformas distintas. Es el porcentaje de los PSM que mapean inequívocamente a la isoforma alternativa.

Cuando calculamos el AP Score para los 150 genes con isorformas alternativas comprobamos que de media obteníamos 515 PSMs discriminantes para la isoforma principal y tan solo ocho PSMs para la segunda isoforma más abundante. Esta observación refuerza la hipótesis de que la mayoría de los genes tienen una isoforma principal, como demostramos en la subsección tercera de esta memoria.

Existen tres grupos de genes sobrerrepresentados en el conjunto *Genes IA*. Quisimos descartar la posibilidad de que las isoformas alternativas detectadas con eventos de *splicing* homólogos, NAGNAG y hnRNPs estuvieran sobrerrepresentadas simplemente porque fueran más fáciles de detectar. Para llevar a cabo estos cálculos estudiamos el número de exones por gen, la probabilidad teórica de detección de isoformas alternativas y el número de PSMs para cada gen.

Los genes del conjunto *Genes IA* se dividieron en ocho grupos en función del evento de *splicing* alternativo:

1. HE: genes con isoformas constituidas a partir de sustituciones de exones homólogos (19).
2. NAGNAG: genes que expresan isoformas alternativas a partir de NAGNAG *splicing* o que tienen inserciones de hasta cuatro residuos (17 genes).
3. hnRNP: genes hnRNP (10).
4. C-sust: genes con isoformas que difieren por sustituciones en el C-terminal.
5. N-sust: genes con isoformas que difieren por sustituciones en el N-terminal.
6. Indels: genes con isoformas que difieren por *indels* de más de cuatro residuos (48).
7. Proteínas diferentes: genes que expresan dos isoformas que no están relacionadas.
8. Complejos: genes con isoformas que difieren a causa de más de un tipo de evento de *splicing* alternativo (16).

Se calculó la puntuación media de cada una de las características para cada uno de los grupos normalizándola en función de los valores medios de todos los genes del conjunto *Genes IA*. Los resultados aparecen en la Tabla 3. Los valores que son más bajos en el conjunto *Genes IA* aparecen marcados en negrita.

	Longitud media de CDS	Media del número de isoformas	Media del número de exones	PSM	Probabilidad de detección
HE	0,54	0,821	0,642	0,848	0,856
NAGNAG	0,784	0,764	0,726	0,495	0,66
hnRNP	0,595	0,826	0,61	0,67	0,362
C-sust	0,836	1,073	0,776	0,391	1,126
N-sust	1,907	1,289	1,367	2,959	1,423
Indels	1,284	1,087	1,277	1,557	0,824
Complejas	1,907	1,26	2,267	1,102	1,752
Proteínas diferentes	0,48	0,768	0,6	0,166	1,783

Tabla 3. Características normalizadas para los ocho grupos de genes detectados con isoformas alternativas en el estudio piloto (*CNIO_proteodata*).

Los genes hnRNP y los que generan isoformas a partir de NAGNAG *splicing* y exones homólogos tienen menos exones, menos variantes y una longitud de secuencia más pequeña que la media. Para estos genes detectamos menos péptidos de media y existe una probabilidad teórica menor de detectar isoformas alternativas. Estas características sugieren que estas isoformas deberían haber sido más difíciles de detectar dentro del conjunto *Genes IA*, sin embargo hemos encontrado muchos más casos de los que esperaríamos en base al número de éstos que están anotados en el genoma.

Se calculó la puntuación AP para todos los genes del conjunto de *Genes IA* y la media de la puntuación AP para cada uno de los ocho grupos. Como podemos ver en la Tabla 3, los genes HE, genes hnRNP y genes con isoformas alternativas generadas a partir de deleciones NAGNAG, tienen unas puntuaciones AP mucho más altas que el resto. La proporción de PSM es mayor tanto para los genes de estos tres grupos como para sus isoformas alternativas.

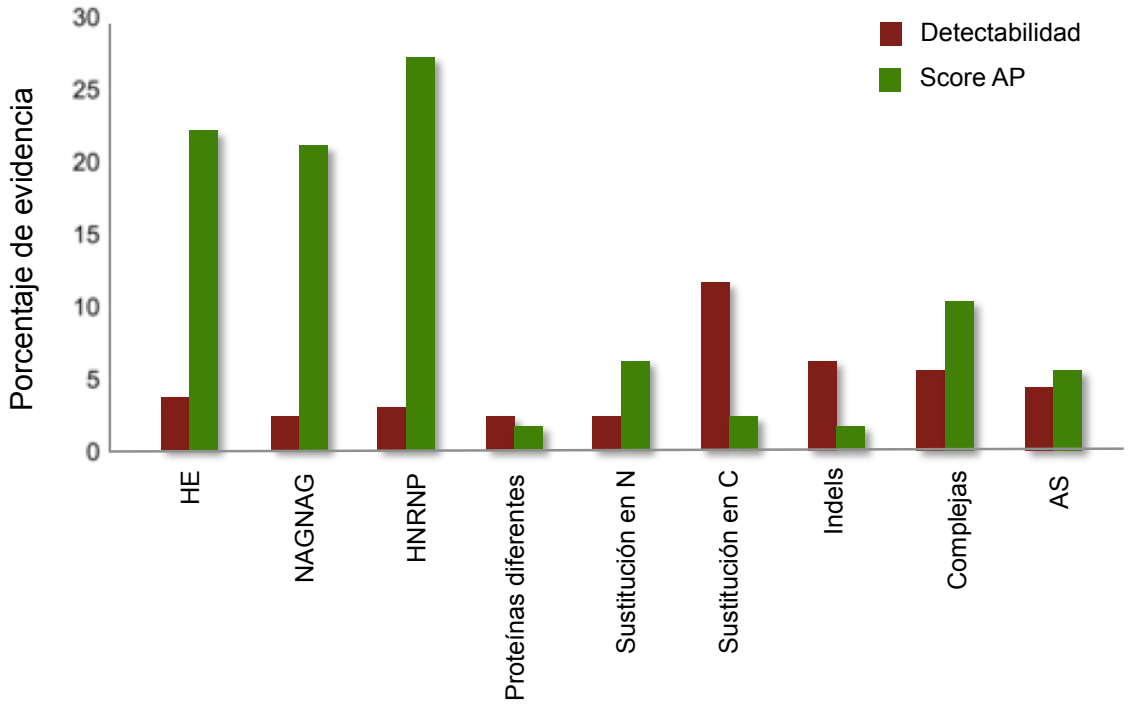


Figura 8. Porcentaje de PSM inequívocos detectados en isoformas alternativas (puntuación AP) frente a la probabilidad teórica de detección para cada uno de los ocho grupos de

isoformas alternativas definidos (análisis realizado sobre el conjunto de datos del estudio piloto *CNIO_proteodata*). Las isoformas alternativas que han sido detectadas en los grupos HE, NAGNAG y hnRNP tienen una puntuación AP muy alta en comparación con la detectabilidad teórica. La detectabilidad se calcula a partir del número de PSMs encontrados en un gen y su probabilidad teórica de detección, normalizando ambas a partir de los valores equivalentes del conjunto *Genes Background*.

3.2.6. Caracterización y significado estadístico de eventos de *splicing* en el estudio ampliado

En el estudio ampliado se encontró evidencia para 282 eventos de *splicing* en 246 genes. Los eventos de *splicing* encontrados con mas frecuencia fueron los *indels* (109 eventos). Los *indels* son también los eventos de *splicing* más comunes en la genoma humano (Mironov, 1999; Mudge *et al.*, 2011). Los segundos más frecuentes fueron las sustituciones homólogas (60 eventos), seguidas de las sustituciones no homólogas en el C-terminal (43 eventos). También se encontraron muchos NAGNAG (39 eventos) y algunas sustituciones en el N-terminal (24 eventos). Por último, obtuvimos casos puntuales de sustituciones internas no homólogas (2 eventos) y eventos generados por proteínas no homólogas (5 eventos). En la Figura 9 podemos ver los tipos de eventos encontrados por experimento individual.

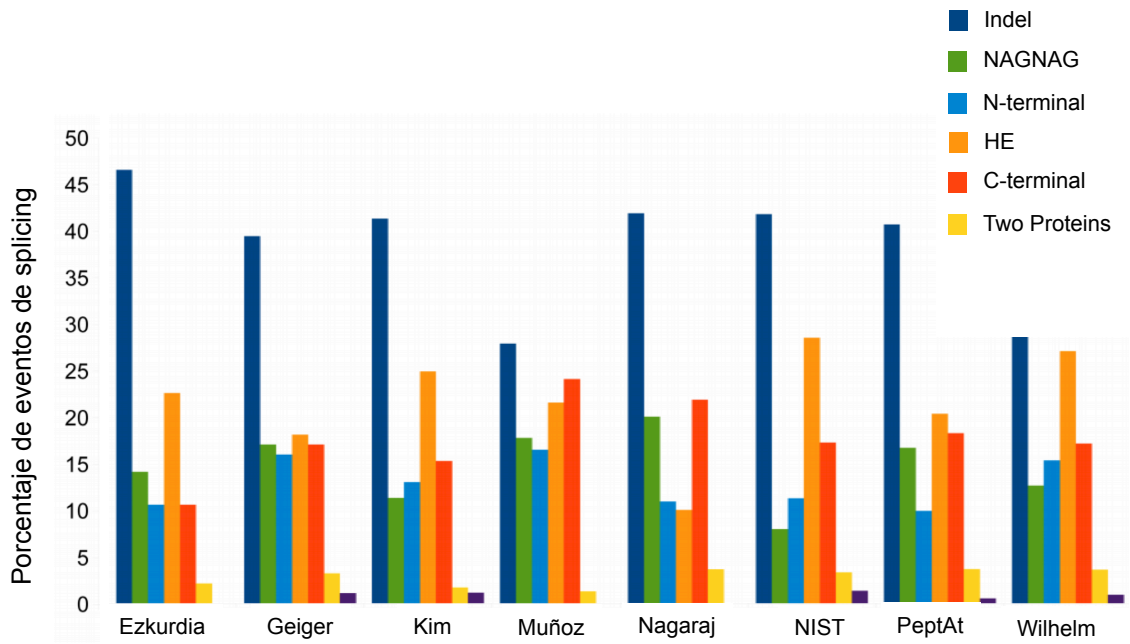


Figura 9. Tipos de *splicing* encontrados en cada uno de los experimentos pertenecientes al conjunto de datos del estudio ampliado (*Eight_proteodata)*.
Porcentajes de cada tipo de evento de splicing encontrado en cada una de las ocho fuentes de datos proteómicos.

Como en el primer estudio, aquí también pudimos constatar que es más fácil encontrar eventos de *splicing* cuanto mayor el número de péptidos detectados por gen. Para ello se calculó la proporción de genes AS en base a los rangos de su abundancia. En la Figura 10 se pueden apreciar las diferencias entre los genes AS y el resto de genes. La facilidad para detectar de los genes AS está relacionada con la abundancia de sus péptidos en los experimentos de proteómica; cuantos más péptidos se recuperen para un gen, más fácil es detectar una de sus isoformas alternativas.

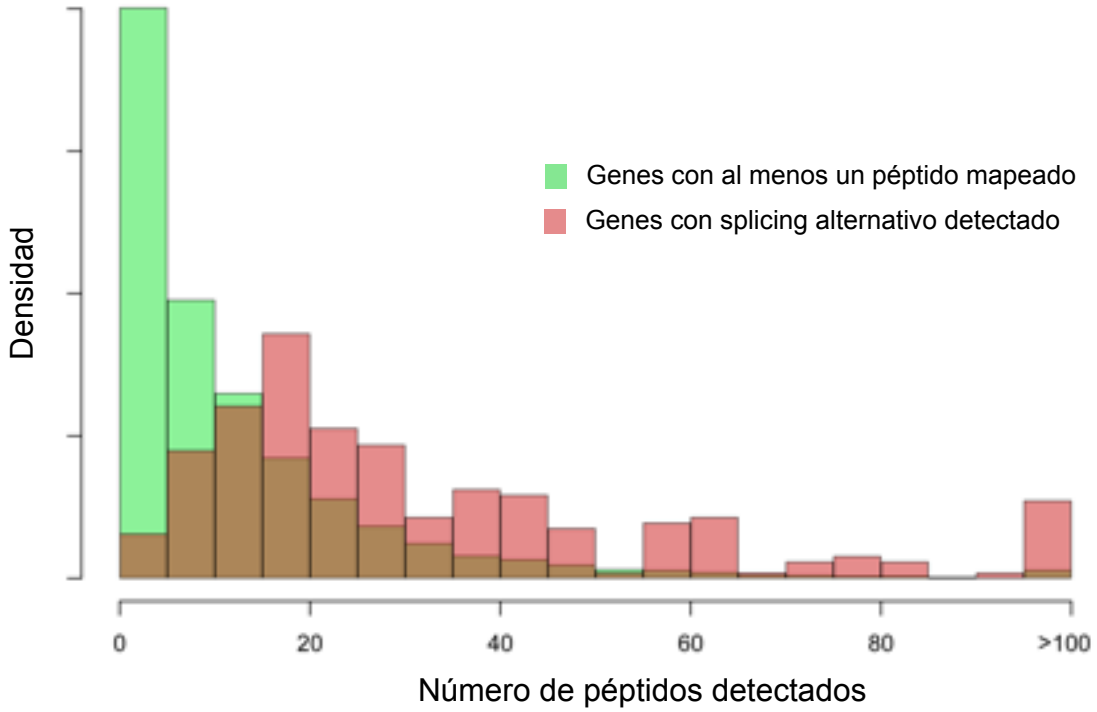


Figura 10. Facilidad de detección y abundancia de genes de *splicing* en el estudio ampliado (*Eight_proteodata*). Histograma de la abundancia de péptidos para los genes con al menos un péptido mapeado (verde) frente a los genes para los que se encuentra *splicing* alternativo (rosa).

También se comprobó si había un sesgo entre los péptidos detectados por gen y el tipo de *splicing* encontrado. Como podemos observar en la Figura 11 no se apreciaron grandes diferencias entre los distintos tipos de *splicing*, salvo para el evento de *splicing* que da lugar a dos proteínas completamente diferentes (*Two*). La prueba de signos de Wilcoxon confirmó, como era de esperar, que la diferencia para este tipo de evento es significativa. De estos resultados podemos inferir que la dificultad de detectar un evento de *splicing* es la misma salvo para aquellos casos en los que el evento dé lugar a proteínas diferentes. Esto implica, que no existe correlación entre la abundancia de eventos HE encontrados y su detectabilidad.

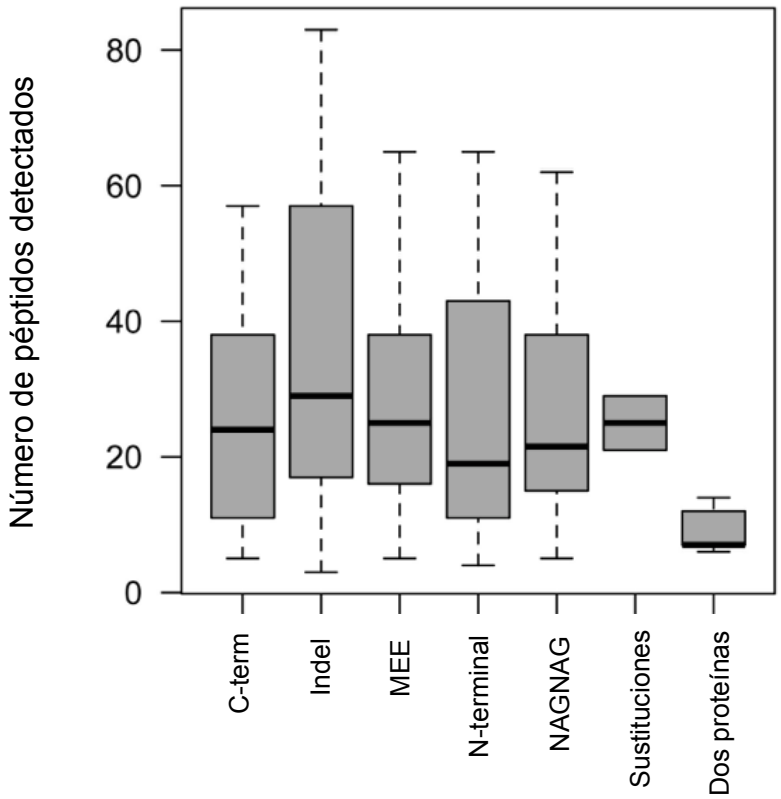


Figura 11. Diagrama de cajas del número de péptidos detectados en base al tipo de evento de *splicing* en el estudio ampliado (*CNIO_proteodata*). En el diagrama de cajas aparencen el máximo y mínimo número de péptidos encontrados para cada uno de los eventos de *splicing* (sustitución en el C-terminal, deleciones, sustituciones de exones homólogos, sustituciones del N-terminal, eventos NAGNAG, sustituciones internas y generación de proteínas diferentes) y sus percentiles.

En el estudio ampliado, al igual que en el estudio anterior, pudimos observar que el número de exones homólogos está sobrerrepresentado. Para comprobar si la proporción de *splicing* HE encontrada está estadísticamente sobrerrepresentada se creó un conjunto de test de 157 genes con eventos HE utilizando BLAST contra GENCODE 20 (ver Métodos y Materiales). Pudimos observar que 33 de los 157 genes de la lista aparecían en nuestros análisis proteómicos. Detectamos un 21% de los genes HE y un 0.01% de eventos de splicing alternativo en GENCODE. El test de Fisher confirmó con p valor cercano a cero que el enriquecimiento de eventos HE es significativo.

Para el caso concreto de HE quisimos descartar también que estos fueran más fáciles de detectar porque tiendan a estar en aquellos genes que más se expresan. Comparamos los péptidos detectados pertenecientes al conjunto de test formado por 157 genes HE y los pertenecientes a los 13.157 genes que están anotados con múltiples isoformas (Figura 12). Calculamos la distribución de frecuencias del número de péptidos encontrados para comparar los dos conjuntos de genes. Con un t-test confirmamos que no había diferencias significativas. Pudimos constatar que el número de eventos HE encontrados no está correlacionado con la expresión del gen.

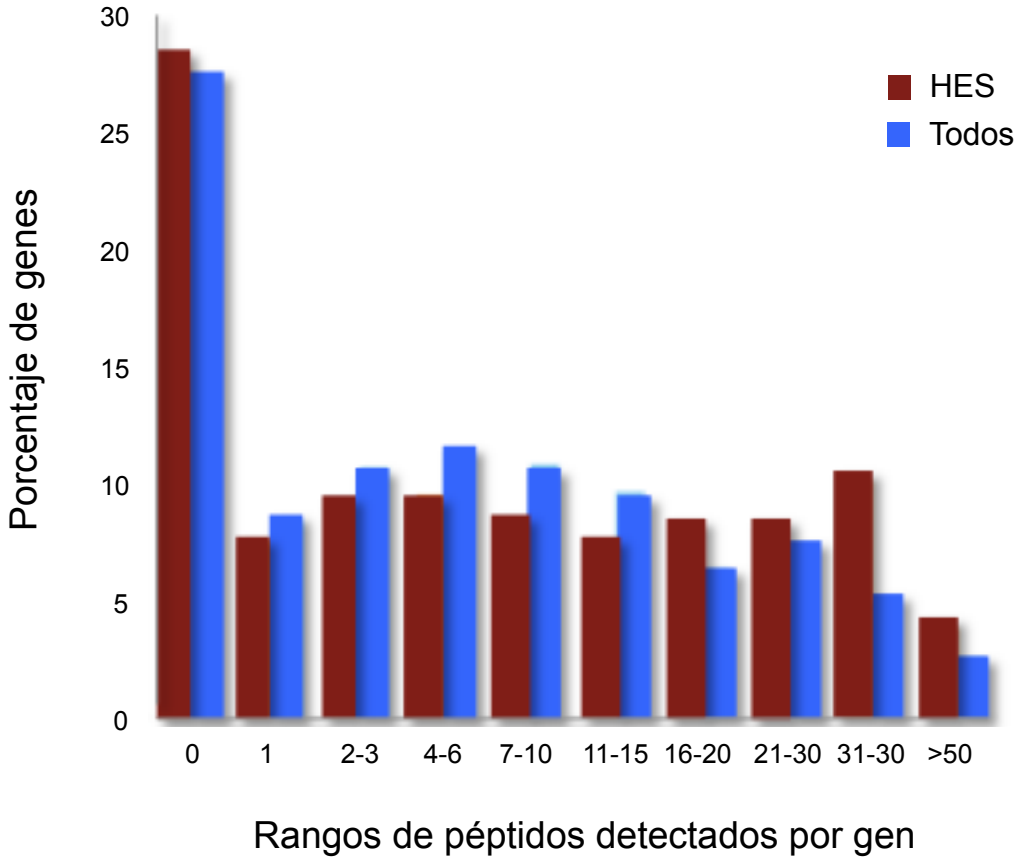


Figura 12. Distribución del numero de péptidos detectados por gen para los que tienen evidencia de exones HE y para los que no la tienen en el estudio ampliado (*Eight_proteodata*). La figura muestra que los genes que tienen evidencia de exones HE

tienen prácticamente la misma distribución de péptidos per gen, que los genes sin evidencia de exones HE.

También se descartó que la proporción del enriquecimiento de eventos HE fuera una artefacto derivado de combinar ocho experimentos proteómicos ya que las proporciones de tipos eventos eran similares. Como podemos ver en la Figura 12 los HE aparecen igualmente representados en todos los experimentos.

Estos resultados confirman que la abundancia de exones hómologos en isoformas alternativas a nivel de proteína es un fenómeno biológico y no está sesgado por los métodos empleados. Una de las razones por las que los HE están sobrerrepresentados podría estar relacionada con la tendencia que hemos observado hacia la identificación de aquellos genes más antiguos y mejor conservados (Ezkurdia *et al.*, 2014a). Esa misma tendencia podría estar ocurriendo también como venimos observando en estos análisis, cuando detectamos isoformas de *splicing* alternativo. En este caso, estaríamos identificando la isoforma alternativa más antigua y mejor conservada, dos características que suelen coincidir con la isoforma generada a partir un evento de HE.

3.2.7. Comparativa de *splicing* alternativo en humano, ratón y mosca del vinagre

Como parte del primer estudio también hicimos una análisis de los datos de mosca y ratón. El análisis de ratón se hizo al igual que el de humano a partir de reanalizar un conjunto de experimentos proteómicos, mientras que el de mosca se realizó a partir de los resultados de dos estudios a gran escala (ver Materiales y Métodos). En la Figura 13. podemos ver las proporciones de los tipos de eventos de splicing más relevantes detectados en los análisis en humano, ratón y mosca

En ratón se detectaron 49 genes con AS. Cuando realizamos la comparación entre humano y ratón vimos que se habían detectado 18 de sus ortólogos humanos, incluyendo 3 de los 4 genes de tropomiosina. De estos ortólogos, 16 compartían el mismo evento de splicing. Diez

de los genes detectados en ratón estaban generados a partir exones homólogos y de éstos seis compartían el mismo evento que sus ortólogos humanos. Siete de los eventos de splicing encontrados en ratón eran NAGNAG y de éstos uno era común al de un ortólogo humano.

También se calculó el marcador AP (porcentaje medio de los PSM que mapean inequívocamente a una isoforma alternativa) para los genes detectados en ratón. Como en el caso de los genes AS detectados en humano, los tipos de evento NAGNAG (27,2%) y HE (39,4%) en ratón eran sustancialmente más altos que el resto de genes (4,2%).

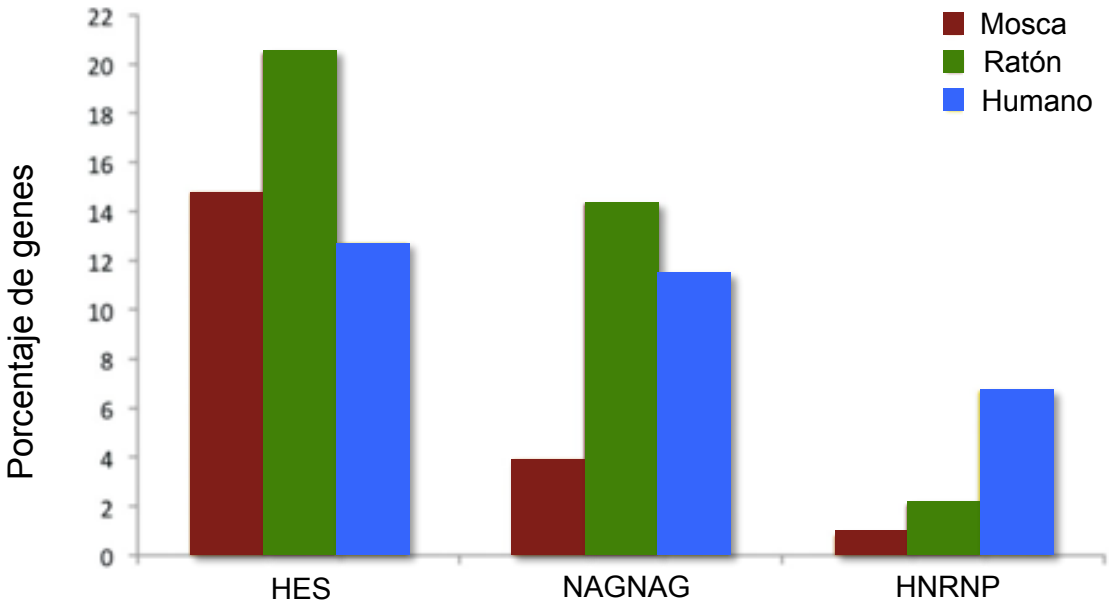


Figura 13. Isoformas alternativas detectadas para humano, ratón y mosca en el estudio piloto (*CNIO_proteodata*).
Porcentajes de los genes detectados con AS que eran de los tipos HE, NAGNAG y HNRNP en mosca, ratón y humano.

En el caso de la mosca se detectaron AS para 130 de los 8.166 genes identificados. 19 de los 130 eventos de splicing encontrados eran HE. La proporción de genes HE es similar a la de ratón y humano aunque se detectó una proporción menor de genes NAGNAG, probablemente porque hay menos anotados en mosca. Los ortólogos de once de los genes AS de mosca aparecieron entre los genes humanos con evidencia de isoformas alternativas. Pero en

ninguno de los casos coincidía el evento de splicing. Es un resultado esperado dado que en la base de datos Flybase solo existen eventos de splicing similares para tres de los 150 genes AS detectados en humano.

En la Figura 14. tenemos un ejemplo del alineamiento de las isoformas alternativas de los genes de tropomiosina en humano y mosca. Ambos genes tienen cuatro grupos de exones homólogos que sean intercambiados, sin embargo, sólo uno aparece en la misma posición en ambas especies.

Figure 14. Exones homólogos anotados en humano y mosca pertenecientes a genes de tropomiosina. ENST00000288398 y ENST00000267996 corresponden a dos isoformas de tropomiosina en humano y CG4898-PL y CG4898-PJ corresponden a dos isoformas del mismo gen en mosca. Los exones homólogos están marcados en colores (azul claro y oscuro para humano y rojo y naranja para mosca), como se puede apreciar en la figura solo hay un par de exones homólogos que solapan y no hay evidencia de que estos sean ortólogos.

Como parte del estudio ampliado se analizaron también experimentos proteómicos adicionales de ratón (ver Materiales y Métodos). En este caso también se pudo observar que las sustituciones homólogas están sobrerrepresentadas en ratón. El 30,1% de los eventos de *splicing* (21 eventos) estaban generados a partir de exones homólogos y todos salvo uno fueron detectados también en sus ortólogos humanos. Este tipo de evento de *splicing* conforma el 60% de los eventos encontrados en ambas especies (Figura 15).

El número de eventos HE detectados en humano y en ratón es extraordinariamente alto. Además como se muestra más adelante, este tipo de evento está conservado en humano y ratón en todos los genes encontrados, mientras que la conservación del resto de eventos anotados es sólo del 19,3%.

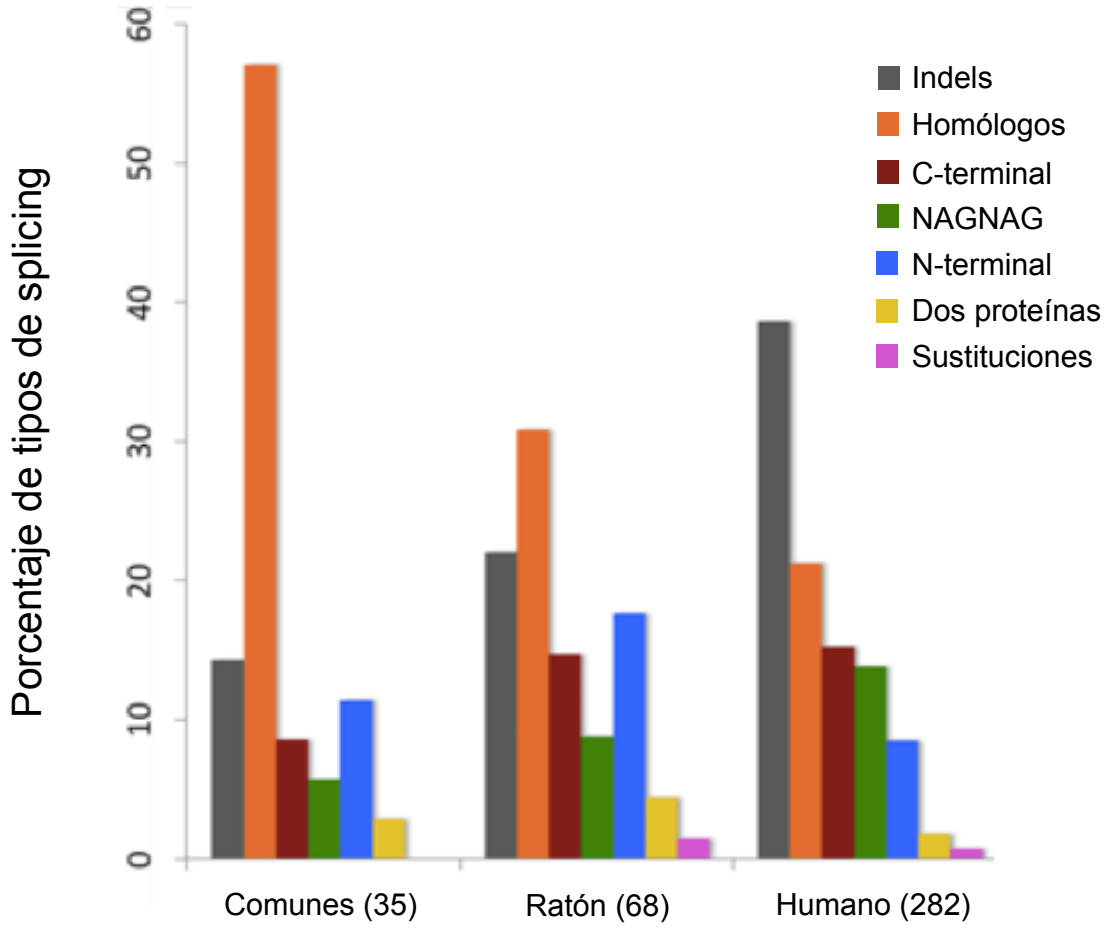


Figura 15. Porcentaje de tipos de *splicing* detectados en el estudio ampliado (*Eight_proteodata*) en humano y ratón. Tanto en humano como en ratón los eventos de splicing detectado más frecuentes son aquellos formados a partir de exones homólogos y deleciones. Como se puede apreciar en la figura cuando tenemos en cuenta aquellos eventos de splicing comunes entre ratón y humano, vemos que aquellos formados a partir de exones homólogos son significativamente altos conformando casi el 60% de los eventos comunes.

3.2.8. Estructuras tridimensionales de las isoformas

No existen muchas estructuras cristalográficas de isoformas alternativas en la base de datos del PDB (Berman *et al.*, 2000). En la literatura sólo habían sido descritas 15 isoformas alternativas con estructura resuelta (Hegyi *et al.*, 2011). Este hecho, probablemente se deba a que en muchos casos los eventos de *splicing* podrían derivar en conformaciones inestables (Melamud & Moult, 2009) e impedir una correcta cristalización.

Cuando comprobamos la existencia de isoformas alternativas en la base de datos del PDB, es requisito que existan estructuras cristalográficas para dos isoformas del mismo gen aunque provengan de especies afines. En el estudio piloto y el estudio ampliado encontramos las estructuras resueltas de dos isoformas detectadas del gen KHK (Figura 16).

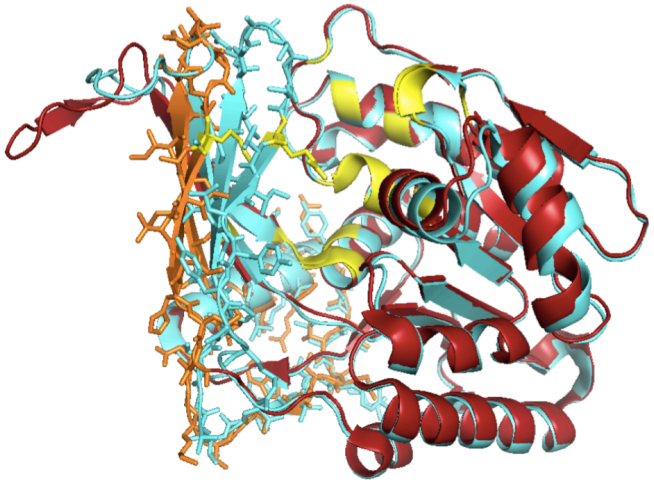


Figura 16. Superposición estructural de las isoformas alternativas de *KHK* detectadas en el estudio ampliado (*Eight_proteodata*).
Los exones homólogos se muestran superpuestos en naranja (PDB 3B3L) y azul (PDB 2HQQ) con las cadenas laterales representadas con *sticks*.

En el segundo estudio se pudieron encontrar diez de los 282 eventos de *splicing* en el PDB (Tabla 4). Nueve de los diez pares encontrados se generan a partir de exones homólogos. Los efectos de estas variantes de exones homólogos sobre la estructura de la proteína tienen una característica en común, no alteran el plegamiento de los dominios en los que se encuentran. Si se tratara de sustituciones no homólogas, la probabilidad de que el plegamiento se viera afectado de alguna manera sería muy alta. Sin embargo, por razones evolutivas, el plegamiento de las estructuras se conserva en secuencias homólogas (Rost, 1999).

Gen	PDB id 1	PDB id 2	Evento de splicing
ACP1	1xww	3n8i	Exones homólogos
CDKN2A	1d9s	1hn3	Dos proteínas
DNM1	4uud	3zvr	Exones homólogos
H2AFY	1zr5	2fxk	Exones homólogos
KHK	3b3l	2hqq	Exones homólogos
MAPK14	1r39	3oht	Exones homólogos
MASP1	4kkd	4igd	Exones homólogos
PFN2	1d1j	2v8c	Exones homólogos
PKM	3srf	3srd	Exones homólogos
SNAP25	1jth	1sfc	Exones homólogos

Tabla 4. Lista de pares de genes detectados en el estudio ampliado (*Eight_proteodata*) con *splicing* alternativo para los que se ha encontrado una estructura en el PDB.

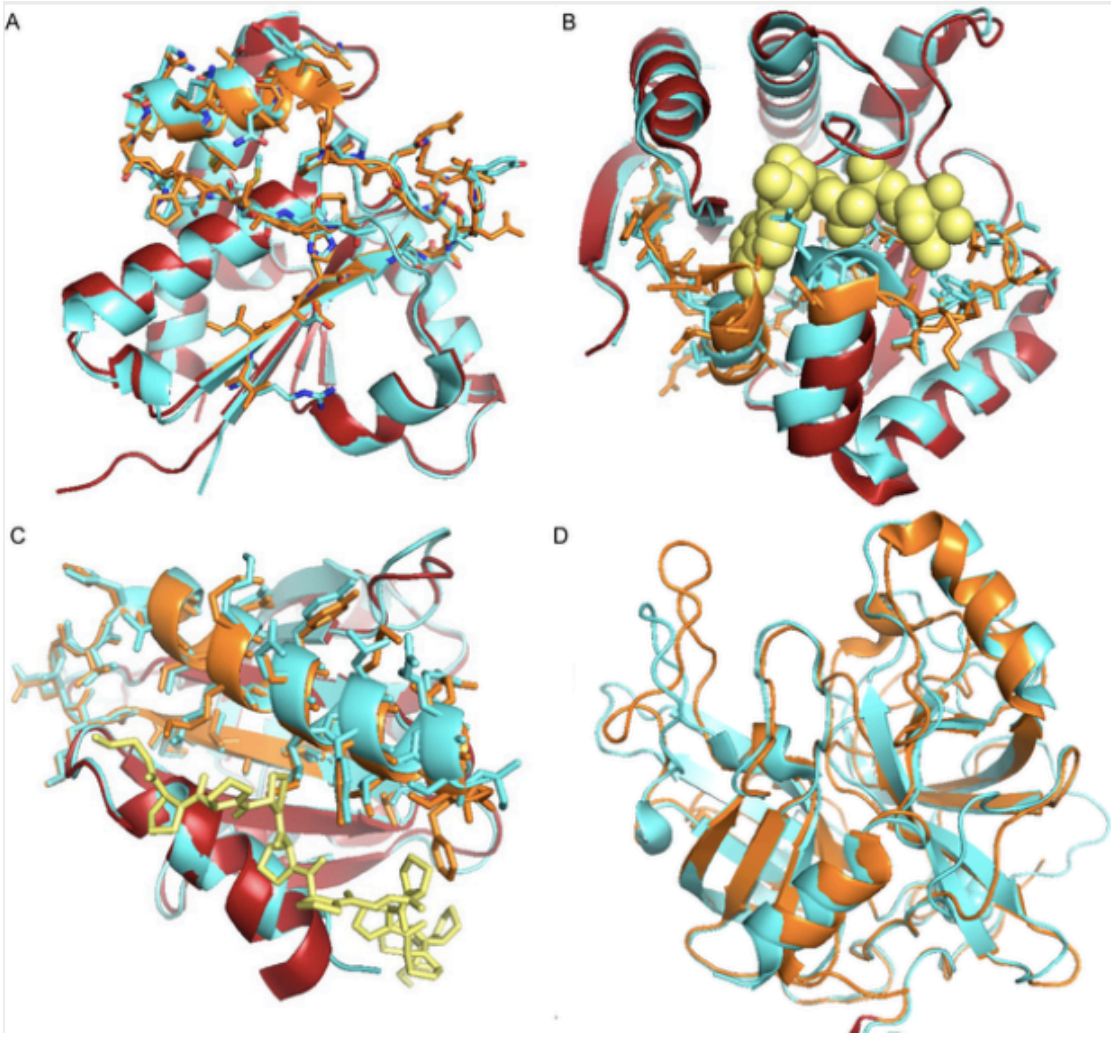


Figura 17. Superposición estructural de isoformas detectadas en el estudio ampliado (*Eight_proteodata*).
Las isoformas aparecen en rojo y azul, las regiones de exones homólogos en naranja y azul. Isoformas generadas a partir de exones homólogos para los genes: A. *ACP1*. B. *H2AFY*. C. *PFN2*. D. *MASP1*.

El plegamiento de proteínas está evolutivamente más conservado que su secuencia de aminoácidos (Rost, 1999), por lo que las variantes de exones homólogos infligen por lo general efectos sutiles en la estructura que pueden incidir sobre la función de las proteínas. Sin embargo deleciones o sustituciones no homologas serían mucho más propensas a modificar o evitar el plegamiento. En la Figura 17 se ilustran varios ejemplos de alineamientos estructurales entre las isoformas de cuatro genes. En el caso del gen *ACP1* la

región afectada forma parte de la superficie de la proteína, mientras que en el caso del gen *H2AFY* incide sobre la unión a sustrato. En el gen *PFN2* la región homóloga podría afectar la unión a sustrato rico en prolina. El gen *MASP1* tiene un dominio de tripsina generado por exones homólogos que comparten menos de un 33% de identidad. Sin embargo la estructura del dominio no se ve afectada.

3.2.9. Análisis funcional

Las relaciones funcionales entre splicing alternativo y la función de proteínas, ha sido investigada por varios grupos a nivel de ARNm, ESTs y microarrays (Modrek *et al.*, 2001b; Pan *et al.*, 2004). Para el primer estudio, se llevo a cabo el análisis funcional para los genes de AS utilizando la herramienta de agrupación de anotaciones funcionales DAVID (Huang *et al.*, 2008) en humano, ratón y mosca. El objetivo de este análisis fue detectar si había un sesgo en las anotaciones funcionales entre los conjuntos *Genes IA* y el conjunto *Genes Background.*

3.2.9.1. Análisis funcional en humanos

Pudimos observar que el conjunto *Genes IA* estaba enriquecido por genes relacionados con el citoesqueleto y la unión a ADN con respecto al conjunto *Genes Background*. Los términos GO significativamente enriquecidos en función del ajuste de p-values por el método de Benjamini utilizado por DAVID fueron: *'actin binding*' (GO:0003779), *'cytoskeletal protein binding*' (GO:0008092), '*structural constituent of muscle*' (GO:0008307), y *'RNA binding*', (GO:0003723).

Cuando comparamos el conjunto *Genes Background* con respecto al conjunto standard definido en DAVID pudimos destacar también un enriquecimiento de términos GO relacionados con actina y citoesqueleto (*'actin filament based process*', GO:0030029, *'cytoskeleton organization*', GO:0007010), regulación de citoesqueleto (‘*regulation of cytoskeleton organization*', GO:0051493) además de términos relacionados con la respuesta

inmune (‘*acute inflammatory response*’, GO:0002526 y ‘*complement activation*’, GO:0006956).

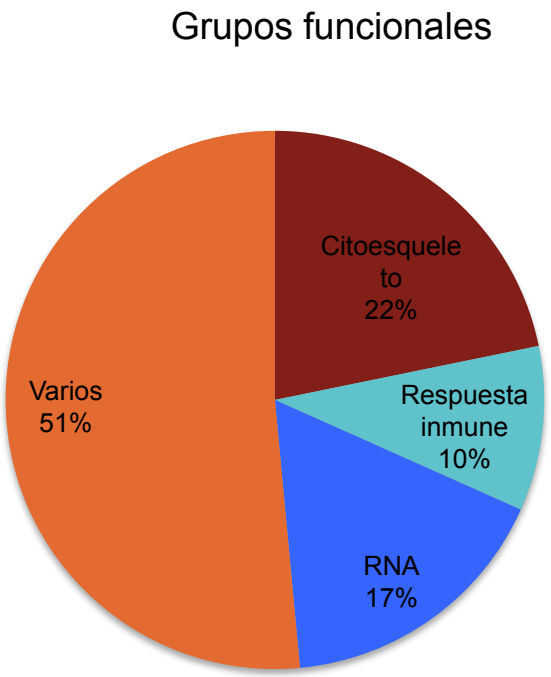


Figura 18. Genes pertenecientes al conjunto *Genes AI* del estudio piloto (*CNIO_proteodata*) agrupados en función a sus anotaciones funcionales.

Los resultados sugieren que el enriquecimiento de terminos GO (Figura 18.) relacionados con citoesqueleto y respuesta inmune, podría deberse en parte a que estas isoformas alternativas serían más fáciles de detectar para estos genes. Para el conjunto de genes relacionados con la respuesta inmune, los niveles de expresión a nivel de transcrito (derivados del Huge Index) y de las proteinas (a partir del número de PSM) fueron bastante mayores.

3.2.9.2. Análisis funcional en ratón y mosca

Los análisis funcionales llevados acabo a partir de los 49 genes para los que se detectó más de una isoforma a nivel de proteína en ratón, mostraron que también estaban enriquecidos con términos relacionados con '*actin cytoskeleton*', sin embargo, no se encontraron terminos sobre representados relacionados con unión a ARN y respuesta inmune como en humano.

En caso de mosca, pudimos observar que los genes que expresan múltiples isoformas alternativas, estaban también enriquecidos por terminos GO relacionados con el citoesqueleto. También se observó un enriquecimiento de terminos GO que se encontraron en ratón y humano como el término relacionado con la ovogénesis (GO:0048477) y la conexión celular "*germline ring canal*" (GO:0045172).

3.2.10. Conservación de dominios Pfam

Resulta de especial interés entender cuáles son los efectos que pueden tener los eventos de *splicing* sobre la función de las proteínas. Se ha demostrado que los dominios están más conservados de lo esperado en las isoformas de *splicing* alternativo (Kriventseva *et al.*, 2003a; Tress *et al.*, 2007a; Severing *et al.*, 2009a). Esto parece sugerir que la selección penaliza de algún modo aquellas isoformas en las que se rompen los dominios Pfam (Tress *et al.*, 2007b), sin embargo, el número de isoformas con los dominios rotos seguiría siendo considerablemente alto.

Con el fin de comprobar si estos dominios estaban también conservados en las isoformas alternativas detectadas en proteómica, utilizamos el programa Pfamscan (Finn *et al.*, 2009) (ver Materiales y Métodos). En base a los alineamientos de las isoformas con los HMM de Pfam, las isoformas se clasificaron como "entera" si todos los dominios se mantienen intactos y como "dañada" si hay uno o más dominios dañados. Para validar los resultados se hicieron los cálculos para todos los grupos de test.

En el estudio piloto pudimos observar que el 45,2% de los dominios de los *Genes background* estaba frente a un 72,3% de dominios dañados en los genes que no se detectaron.

Es conveniente resaltar que para este cálculo, a diferencia del que se muestra a continuación para el estudio ampliado, no se consideraron los eventos de *splicing*. Simplemente se sumaron los porcentajes de dominios rotos en cada uno de los conjuntos. Hay que tener en cuenta que un número bastante alto de los dominios del genoma humano aparecen como rotos, en parte porque no siempre están bien definidos. Sin embargo, en el gráfico se puede apreciar con claridad el efecto del *splicing* sobre los dominios. El porcentaje de dominios rotos para el conjunto de isoformas no detectadas en *Genes AI* es considerablemente mayor al de las isoformas principales detectadas en el conjunto *Genes AI*.

En el estudio ampliado pudimos comprobar mucho mejor el efecto que ejercen los 282 eventos ISE encontrados sobre los dominios Pfam refinando nuestra estrategia de análisis. Los dominios de PfamA se mapearon a todas las isoformas con el programa Pfamscan. En este caso consideramos el dominio roto si el evento de splicing implicaba la perdida o ganancia de cinco o más residuos.

Aquí pudimos comprobar de manera mucho más clara que los dominios Pfam se mantienen intactos en el 84,4% de eventos de *splicing* encontrados. En el 7,1% de los casos se pierden uno o más dominios y en el 6,7% de los casos el dominio se rompe.

Con el objetivo de validar los resultados se calculó el efecto que ejercen todos los eventos de *splicing* de GENCODE 20. Estos cálculos se llevaron a cabo tomando como referencia la isoforma principal predicha por APPRIS (Rodriguez *et al.*, 2012) para cada gen. A partir de estas isoformas principales se generaron los alineamientos con todas las isoformas alternativas y se mapearon las anotaciones de PfamA a todos los eventos de *splicing* (ver Materiales y Métodos).

También se generaron cuatro subconjuntos de GENCODE 20 para descartar sesgos estadísticos: "Detected", que incluye los 12.716 genes para los que se detectan péptidos,

"Min 20" y "Min 50", formados por 2.271 y 385 genes que son detectados con al menos 20 y 50 péptidos respectivamente y "ISE genes" formado por los 246 genes para los que detectamos isoformas alternativas.

Cuando se tienen en cuenta todas las isoformas alternativas de GENCODE 20 (45.346 eventos de *splicing*), comprobamos que el dominio Pfam se rompe en el 37,3% (16.937 eventos) de las ocasiones y supone la pérdida de uno o más dominios en el 32% (14.766) de los casos. Como podemos observar en la Figura 19, los resultados obtenidos para los cuatro subconjuntos de test son muy similares.

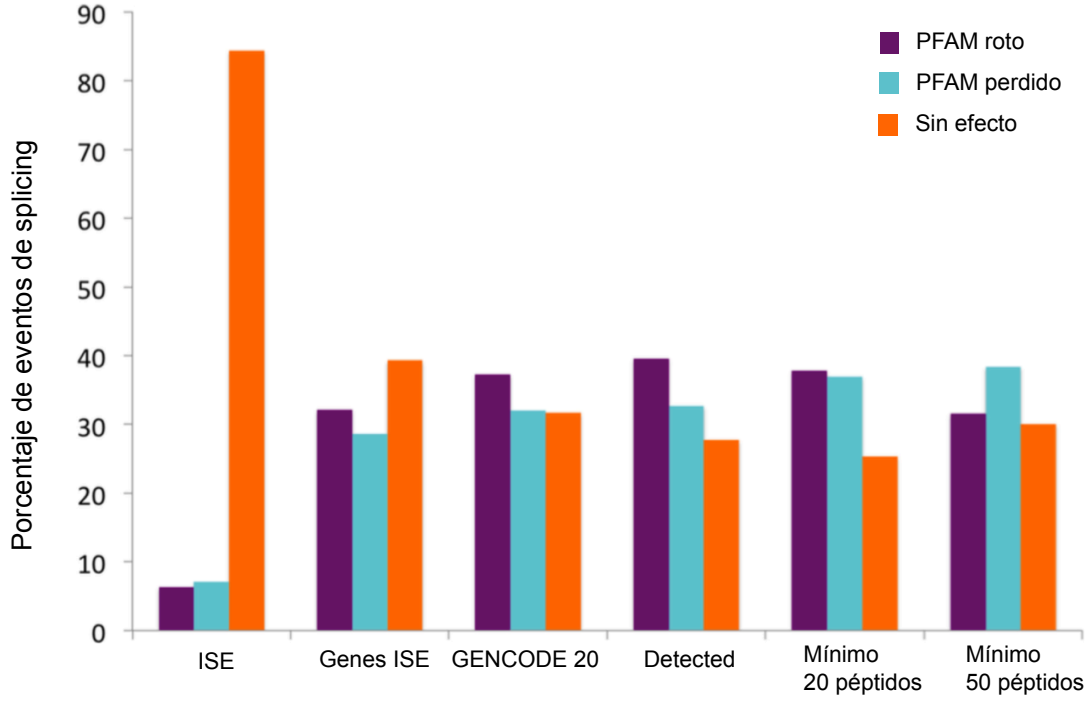


Figura 19. Efecto de los eventos de splicing en dominios funcionales Pfam.
El porcentaje de eventos que producirían isoformas alternativas con dominios Pfam rotos (PFAM roto), en los que se perdería un dominio completo (PFAM perdido) y en los que la composición de dominios de la isoforma alternativa permanecería intacta (Sin efecto). Se muestran los porcentajes para seis tipos de eventos, todos los eventos anotados a nivel de transcrito para todos los genes codificantes anotados en GENCODE 20 (GENCODE20), todos los eventos anotados a nivel de transcrito para los genes identificados en el análisis de

proteómica (*Detected*), todos los eventos anotados a nivel de transcrito para los genes identificados con al menos 20 péptidos (Mínimo 20 péptidos) y para los genes identificados con al menos 50 péptidos (Mínimo 50 péptidos), todos los eventos de *splicing* anotados a nivel de transcrito en los 246 genes para los que se identifican eventos de *splicing* (Genes ISE) y finalmente los 282 eventos de *splicing* para los que identificamos péptidos (ISE).

Estos resultados confirman que la gran mayoría de isoformas alternativas con dominios Pfam rotos no se expresan en cantidades suficientes para ser detectadas en experimentos de proteómica. Además, refuerza la hipótesis de que existe una presión selectiva hacia la conservación funcional de los dominios a nivel de traducción.

3.2.11. Expresión específica de tejido

Muchos estudios sugieren que el *splicing* alternativo es sobre todo un fenómeno específico de tejido o línea celular (Xu *et al.*, 2002; Buljan *et al.*, 2012). La especificidad del *splicing* alternativo se ha podido constatar en experimentos de transcriptómica pero resulta más complicado hacerlo en experimentos de proteómica. En primer lugar, no hay muchos experimentos específicos de tejido o línea celular que abarquen toda la diversidad de los mismos. Además la cobertura de las identificaciones está limitada por varios factores como las características técnicas del espectrómetro de masas en términos de poder de resolución, sensibilidad y exactitud en la medida de masa y a la complejidad de las muestras. Por último, la falta de reproducibilidad que se ve agravada cuando las concentraciones son mínimas como podría ser el caso en la expresión de isoformas alternativas. Quizás por todos estos problemas, hasta ahora no se han publicado estudios en los que se haya demostrado la existencia de *splicing* específico de tejido a nivel de proteómica. Pero además de todas estas consideraciones técnicas, conviene recordar que para lograr detectar según que isoforma alternativa, estamos supeditados, en el peor de los casos, a la posibilidad de encontrar un solo péptido que la distinga inequívocamente.

Afortunadamente, para este trabajo se pudo contar con datos proteómicos específicos de tejidos (Kim et al., 2014), que además habían sido confirmados con hasta tres réplicas, lo que permitió hacer un análisis específico de tejido.

A pesar de esta cantidad de datos experimentales solamente se encontró evidencia de isoformas específicas de tejido para 14 genes que consideráramos altamente fiables (debía poder detectarse en al menos dos de las tres réplicas). De estos genes, nueve se generan a partir de *splicing* de exones homólogos. Al menos uno de cada par de estas isoformas alternativas se encontraron en tejidos de corazón (de adulto o de feto), o en tejido de córtex de adulto o en cerebro de feto.

El tejido en el que encontramos más isoformas específicas es el de corazón. Siete de las nueve isoformas están relacionadas con el citoesqueleto y muchas codifican proteínas en la línea Z del sarcómero. En cerebro encontramos cinco isoformas específicas de tejido, muchas de ellas relacionadas también con el citoesqueleto.

La ausencia de más isoformas alternativas específicas de tejido no implica necesariamente que estas no estén expresadas en dichos tejidos por todas las razones que hemos expuesto. Sin embargo, tenemos que tener en cuenta también indicios que apuntan a que quizás, el splicing específico de tejido pudiera no ser tan común como se creía (Melé *et al.*, 2015). El consorcio GTex (Melé *et al.*, 2015) sugiere que al menos una parte de los transcritos alternativos detectados por RNAseq podrían ser simplemente ruido biológico. El esclarecimiento de este fenómeno tiene una gran importancia para entender mejor la maquinaria celular.

3.3. Análisis de isoformas principales a nivel de proteína

La existencia de *splicing* de mRNA ha sido demostrada a partir de diferentes técnicas como EST, cDNA (Harrow *et al.*, 2006) y datos de *microarrays* (Johnson *et al.*, 2003). Las bases de datos de anotación manuales como GENCODE evidencian la existencia de múltiples

transcritos alternativos de mRNA. La versión 20 de GENCODE tenía mas de 20,000 genes codificantes anotados y alrededor de 93,000 transcritos codificantes.

Existe controversia a la hora de determinar cuál es el transcrito que se expresa a nivel de gen en diferentes tejidos. Estudios basados en ESTs (Taneri *et al.*, 2011) sobre 13 tejidos predicen un transcrito dominante por gen; sin embargo, otros estudios basados en RNAseq (Djebali *et al.*, 2012; Gonzàlez-Porta *et al.*, 2013) estiman que una proporción de los genes tendría más de un transcrito dominante específico de tejido.

Hay estimaciones que sugieren que splicing alternativo se podría dar en el 95% de los genes humanos que tienen varios exones (Pan *et al.*, 2008; Wang *et al.*, 2008). El inventario de proteínas posibles aumentaría considerablemente en el supuesto de que estos transcritos fueran traducidos en proteínas (Smith & Valcárcel, 2000).

Todavía no se han publicado estudios que investiguen cuál es la isoforma principal expresada a nivel de proteína en diferentes tejidos. En este sentido, nos propusimos averiguar cuáles son los transcritos que se expresan a nivel de proteína en diferentes tejidos. Así mismo, estudiamos la relación de expresión de los transcritos a nivel de gen y de proteína.

En los estudios anteriores se pudo observar que siempre encontrábamos una isoforma con un número de PSMs considerablemente mayor al resto de isoformas detectadas para el mismo gen. Esto reforzó la hipótesis de que las isoformas alternativas podrían estar siendo traducidas de forma ocasional y de que habría una isoforma que estaría siempre más expresada. En el análisis que se expone a continuación se profundizó en el análisis de isoformas principales a nivel de proteína.

El hecho de detectar isoformas alternativas para tan sólo un 1,2% de los genes (246 genes) detectados en proteómica supone una ventaja a la hora de determinar cuál es la isoforma principal, puesto que un recuento simple de péptidos que mapean a una isoforma inequívocamente podría ser suficiente para poder determinar cuál es la isoforma principal. Para llevar a cabo este estudio se utilizaron el conjunto de datos *Eight_proteodata.* Primero,

se descartaron los 3,977 genes anotados con una sola isoforma codificante. Para el resto de genes se consideró aquella isoforma con más péptidos asignados. De esta manera, se detectaron 5.011 genes para los que fuimos capaces de distinguir la isoforma principal. Para 3.703 genes no se logró distinguir la isoforma principal, al no haber suficientes péptidos que pudieran distinguir entre isoformas. También hallamos 25 genes para los que el número de péptidos discriminantes entre isoformas era el mismo, todos ellos dentro del conjunto de 246 genes con isoformas alternativas.

Con el propósito de determinar si la asignación de isoformas tenía sentido biológico, se compararon con los resultados publicados recientemente en el estudio de Human BodyMap (Gonzàlez-Porta *et al.*, 2013), las variantes CCDS *(Consensus Coding Sequences*) (Pruitt *et al.*, 2009) y el método de predicción de isoformas principales APPRIS.

3.3.1. Comparativa con RNASEQ

En un trabajo reciente (Gonzàlez-Porta *et al.*, 2013), se analizaron los datos de RNAseq en múltiples tejidos y líneas celulares. Los autores asignaron un transcrito principal basándose en la condición de que el transcrito tuviera una expresión cinco veces mayor para todos los tejidos o para todas las líneas celulares. De esta manera encontraron 4.199 genes de líneas celulares y 5.228 genes de tejidos que cumplían esta condición. La base de datos de referencia que utilizaron fue la versión de GENCODE 11. Debido a las diferencias entre esta versión de GENCODE y la empleada en nuestro estudio (GENCODE 20), para poder comparar nuestros resultados fue necesario descartar todos los transcritos y genes con identificadores distintos. También se excluyeron todos aquellos genes para los que se había asignado la variante principal a un transcrito no codificante. El 20% de los transcritos dominantes encontrados por Human BodyMap fueron asignados a transcritos no codificantes. Los autores sugieren que muchos de los transcritos expresados podrían estar jugando una función regulatoria y no ser traducidos. Este proceso nos permitió comparar 1.038 genes encontrados en tejidos y 762 en líneas celulares. El solapamiento entre las isoformas principales proteómicas y las sugeridas por el estudio de Human BodyMap basado en RNAseq es de un 72,2% en tejidos y un 81,6% en líneas celulares (Tabla 5).

Las diferencias de expresión entre proteómica y RNAseq podrían estar también relacionadas con el recambio proteico (Ning & Nesvizhskii, 2010; Sheynkman *et al.*, 2013), la edición post-transcripcional (Farajollahi & Maas, 2010) o la eficiencia de traducción de proteína (Vogel & Marcotte, 2012).

3.3.2. Comparativa con APPRIS y CCDS

En la base de datos de variantes CCDS había 13.297 genes anotados que aparecían en la base de datos GENCODE 20 utilizada en nuestro estudio. Así, pudimos comparar 3.331 genes de los 5,011 genes para los que se tenía evidencia proteómica y estaban también anotados en CCDS. Se comprobó que la isoforma principal era la misma para el 98.6% de los casos (Tabla 5).

En el momento del estudio, en APPRIS existían 15.172 genes de GENCODE 20 anotados para los que se pudieron comparar 4.186 genes obteniendo una coincidencia del 97,8% con nuestros resultados (Tabla 5.). Existen 3.015 genes para los que existe evidencia proteómica, están anotados en CCDS y además aparecen en APPRIS. En este conjunto, las isoformas principales que coinciden aumentan hasta en un 99,37% con CCDS y hasta en un 99,5% con APPRIS.

	a) genes	b) comparables	c) discrepancia	d) % consenso	e) % discrepancia
CCDS unique	13297	3331	46	98,6	1,4
isoforma principal APPRIS	15172	4186	93	97,8	2,2
isoforma más larga	20462	5011	520	89,6	10,4
Celulas HBM	4199	762	140	81,6	18,4
Tejidos HBM	5228	1038	237	77,2	22,8
***reads* sin filtrar**	15950	4618	656	85,8	14,2
***reads* dobles**	2465	1018	62	95,4	4,6

Tabla 5. Comparativa de isoformas principales

a) Número de genes para los que se pudo determinar la isoforma de referencia. b) Numero de genes para los que se pudieron hacer las comparaciones con la isoforma principal de proteómica. c) Número de genes para los que hay un desacuerdo entre la isoforma referencia y la isoforma proteómica. d) Porcentaje de genes para los que hay acuerdo entre la isoforma referencia y la isoforma proteómica principal. e) Porcentaje de genes para los que no hay acuerdo entre la isoforma referencia y la isoforma proteómica principal.

Los resultados confirman una coincidencia sobresaliente con las bases de datos CCDS, creada a partir de la curación manual de evidencia genómica, y APPRIS, basado en la conservación de estructura y función.

A la hora de seleccionar la isoforma principal muchas bases de datos siguen optando por sugerir la variante más larga. Cuando se comparó la isoforma más larga con el conjunto de 5.011 genes, obtuvimos unos resultados peores, con un 89.6% de asignaciones coincidentes.

Ante estos resultados, investigamos cuáles podrían ser las causas de discrepancia entre las isoformas principales con evidencia experimental proteómica y las sugeridas por CCDS y APPRIS. Una exploración manual de estos casos nos mostró que en muchas ocasiones se deben a problemas relacionados con la definición del modelo génico. Cabe resaltar que para aquellos casos en los que la isoforma principal se puede asignar tanto por proteómica, como por APPRIS y CCDS, el acuerdo de estos tres métodos ortogonales era del 99%.

4. Discusión

El trabajo presentado en esta tesis tiene como objetivo el desarrollo de un flujo de trabajo para el análisis de datos de proteómica a gran escala, la verificación de anotaciones genómicas a nivel de proteína, la caracterización de isoformas de *splicing* alternativo y el estudio de las isoformas principales.

4.1. Complejidad del proteoma y limitaciones en la detectabilidad de péptidos

Existen varios factores que limitan las probabilidades de detectar un péptido en un experimento de proteómica. Entre los factores técnicos relacionados con la tecnología de espectrometría de masas está el de la sensibilidad del espectrómetro de masas. Las proteínas tienen que expresarse en cantidades suficientes para poder ser detectadas. Por otro lado, algunos péptidos ionizan peor que otros haciendo más compleja su detección. Existen también otros factores relacionados con la estructura de la proteína que influyen en la probabilidad de detección de algunos de sus péptidos, como en el caso de las proteínas formadas por múltiples hélices transmembrana, más difíciles de detectar (Ezkurdia *et al.*, 2014a).

Hay que tener también presente que la no detección no implica que el gen no se esté expresando; las proteínas podrían estar presentes tan solo en ciertos tejidos, activarse solo en casos muy concretos (Lane et al., 2014) o en etapas de desarrollo específicas. Probablemente existan algunas proteínas e isoformas para las que no hayamos encontrado evidencia por ser específicas de tejido o línea celular, a pesar de que hemos incluido datos de proteómica en los que se han utilizado tejidos embrionarios y células madre pluripotentes (Munoz *et al.*, 2011; Kim *et al.*, 2014). Algunas proteínas además tienen una vida media muy corta como en el caso de las proteínas de la familia HOX (Lane et al., 2014), lo que complica su detección.

Por otro lado cabe resaltar la complejidad que implica la validación de isoformas, supeditada a las diferencias a nivel de secuencia que permiten la detección de los péptidos capaces de distinguirlas.

Es importante resaltar la importancia de la utilización de métodos de validación apropiados cuando se analizan datos proteómicos a gran escala (Reiter *et al.*, 2009; Savitski *et al.*, 2015). En dos trabajos recientes (Ezkurdia *et al.*, 2014b; 2015a) demostramos las consecuencias de la aplicación inadecuada de estos métodos (Kim et al., 2014; Wilhelm et al., 2014) para los que se genera un número elevado de falsos positivos.

Para esta memoria se han generado dos conjuntos de péptidos a partir de un análisis de validación conservador. A pesar de que esta aproximación probablemente haya reducido la sensibilidad de detección de algunas isoformas alternativas, consideramos que era importante anteponer la calidad a la cantidad con el objetivo de caracterizar el *splicing* alternativo a nivel de proteína de la manera más verosímil. Y de este modo, producir un conjunto de isoformas fidedigno que pueda utilizarse como referencia para otros trabajos.

4.2. Aportaciones a las anotaciones genómicas

La detección de péptidos en diversas fuentes de datos proteómicos a gran escala, utilizando tres versiones diferentes de la base de datos para humano de GENCODE, ha permitido confirmar, completar y corregir muchas anotaciones genómicas. Con el estudio hemos logrado detectar evidencia de transcritos candidatos a NMD, transcritos anotados como pseudogenes y transcritos anotados como "putative" y "novel". Además, con este método se ha caracterizado el *splicing* alternativo a nivel de proteína y confirmado de manera fehaciente cuales son las isoformas principales a nivel de proteína.

La detección de cuatro transcritos susceptibles de ser degradados por mediación de mutación terminadora, sugiere que estos mecanismos podrían no ser del todo eficientes o que no se entiende completamente su funcionamiento. Recientemente se han propuesto mecanismos de regulación que podrían interferir en el proceso de NMD. El microARN (miR-128) específico

del cerebro, se une a la ARN helicasa UPF1 reprimiendo el sistema NMD (Bruno et al., 2011). Análisis específicos en proteogenómica podrían ayudar en la detección de transcritos susceptibles a NMD y confirmar diferentes hipótesis sobre el funcionamiento de las mismas.

La espectrometría de masas es un método idóneo para la identificación de proteínas y por lo tanto una herramienta muy útil para la caracterización de anotaciones genómicas siempre y cuando estas se expresen en cantidades suficientes para ser detectadas por el espectrómetro de masas. El flujo de análisis de datos proteómicos a gran escala utilizado en esta memoria nos ha permitido profundizar en la caracterización proteogenómica humana en otro estudio publicado recientemente (Ezkurdia et al., 2014a). En este trabajo hemos establecido nuevas relaciones entre la detectabilidad a nivel de proteína, las divisiones filogenéticas y la conservación entre especies, e identificado un set de 2001 genes humanos que podrían estar anotados erróneamente como codificantes. Estos genes se enviaron a los anotadores y GENCODE ya ha descartado más de mil de la base de datos a partir de revisiones manuales. Puesto que el genoma humano sirve de referencia en la anotación del resto de vertebrados, la revisión de estos genes supone un impacto enorme tanto en especies homólogas como en especies en las que las validaciones se hacen de manera automática a partir de la humana.

4.3. *Splicing* alternativo a nivel de proteína

A partir de los datos de PeptideAtlas y GPM hemos detectado 150 genes que expresan varias isoformas de *splicing* alternativo. Posteriormente hemos repetido el estudio piloto utilizando ocho fuentes diferentes de datos proteómicos a mayor escala y ampliando el número de genes detectados a 246.

Muchos autores han sugerido que el *splicing* alternativo estaría contribuyendo a un aumento de complejidad funcional en las células (Modrek & Lee, 2003), basándose en la evidencia que existe a nivel de transcrito. Si partimos de esta hipótesis, se podría esperar encontrar un número de isoformas alternativas de *splicing* sustancialmente mayor, teniendo en cuenta la cantidad de muestras que han sido analizadas en este trabajo.

Sin embargo, el número de isoformas alternativas que validamos en nuestros estudios es pequeño, un orden de magnitud por debajo del estimado cuando simulamos *in silico* el número de isoformas que obtendríamos si la probabilidad de detección de cualquier isoforma fuera la misma. Al margen de todos los factores que pudieran estar influyendo a la hora de detectar isoformas a nivel de proteína, los resultados obtenidos sugieren que la diversidad de las mismas es mucho menor que la estimada previamente.

Cuando analizamos el conjunto de isoformas alternativas detectadas, podemos constatar que muchas de ellas han sido bien caracterizadas en la literatura, aparecen implicadas en procesos celulares, están conservadas en ratón o han sido generadas a partir de pequeños cambios en sus secuencias dando lugar a cambios funcionales sutiles.

Estos resultados sugieren que a pesar de que muchas variantes de *splicing* alternativo aparecen expresadas a nivel de transcrito, tan solo un porcentaje es traducido a proteína en cantidades suficientes para ser detectado. Esta hipótesis concuerda con la de otros estudios que sugieren que la expansión de diversidad funcional estaría limitada (Tress *et al.*, 2007b; Severing *et al.*, 2009b) o que muchas de estas isoformas alternativas podrían ser el resultado de errores en el proceso de *splicing* (Melamud & Moult, 2009; Pickrell *et al.*, 2010).

4.4. Sobrerrepresentación de exones homólogos

Una de las características más relevantes a la hora de caracterizar los patrones de *splicing* alternativo detectados es la sobrerrepresentación de isoformas alternativas formadas a partir de la sustitución de exones homólogos. Este patrón de *splicing* aparece en un 13% de las isoformas alternativas detectadas para el primer conjunto de datos proteómicos analizados y en un 20% de las isoformas alternativas detectadas para el segundo conjunto de datos, muy por encima del 0,01% que esperaríamos encontrar cuando hacemos una simulación *in silico*. La sustitución de exones homólogos también está presente en el 60% de los eventos de *splicing* ortólogos encontrados entre humano y ratón.

Cuando hemos estudiado la conservación de este patrón de *splicing* para todas las isoformas alternativas detectadas en el segundo estudio, hemos visto que estas tendrían un ancestro común en los primeros vertebrados con mandíbula hace más de 460 millones de años. Estas observaciones abren nuevas vías de investigación que permiten profundizar en la relación de la duplicación y el *splicing* alternativo (Abascal *et al.*, 2015b).

También hemos podido comprobar que muchas de las isoformas alternativas detectadas con este patrón o las de sus homólogos tienen su estructura resuelta en la base de datos PDB, y que los cambios de exones homólogos no influyen en el plegamiento de estas estructuras.

Estos hallazgos consolidan la generación de isoformas alternativas a partir de la sustitución de exones homólogos como un mecanismo idóneo que permite cambios sutiles de función sin alterar el plegamiento de la proteína.

4.5. Conservación funcional

Está demostrado que los dominios Pfam de las isoformas de *splicing* alternativo se rompen en menor medida a la que esperaríamos al azar (Kriventseva *et al.*, 2003b; Tress *et al.*, 2007b). Sin embargo, a diferencia de las variantes de *splicing* anotadas en GENCODE, propensas a sufrir una rotura de sus dominios Pfam, la inmensa mayoría de las isoformas alternativas detectadas a nivel de proteína en los dos estudios mostrados conservan intactos sus dominios Pfam.

En el primer estudio podemos comprobar también que cuando analizamos 75 isoformas alternativas que están conservadas en ratón, los dominios Pfam aparecen conservados en prácticamente todos los casos. Esto sugiere que la conservación funcional y estructural de dominios juega un papel crucial en el desarrollo de las isoformas de *splicing* alternativo.

Esta tendencia hacia los cambios sutiles en estructura y función entre las isoformas detectadas en experimentos de proteómica sugiere que podrían existir mecanismos de regulación (Schubert *et al.*, 2000; Maximilian Wei-Lin Popp, 2013; Lykke-Andersen & Bennett, 2014).

Estos impedirían la traducción de aquellos transcritos que pudieran llegar a dañar la célula en el caso de tener sus funciones truncadas.

4.6. Ribonucleoproteínas Heterogéneas-Nucleares

En este estudio hemos visto una sobrerrepresentación de los genes de ribonucleoproteínas heterogéneas-nucleares (hnRNP). En el estudio piloto hemos encontrado evidencia de isoformas alternativas para 10 de los 26 genes hnRNP anotados con isoformas alternativas en GENCODE, además de una mayor proporción de péptidos que mapean a estas isoformas. Este enriquecimiento de genes hnRNP ha sido descrito también en trabajos previos (Tanner *et al.*, 2007b).

El hecho de que hayamos encontrado una evidencia mayor en la traducción a proteína de estas isoformas hnRNP es especialmente interesante, dado que estos genes están directamente relacionados con los procesos de generación de transcritos mRNA alternativos (Dreyfuss, Kim, & Kataoka, 2002; Martinez-Contreras et al., 2007; Matlin, Clark, & Smith, 2005; Venables et al., 2008). Estos genes están asociados con los precursores mRNA en el núcleo, e interfieren en el procesado del ARN y la selección de exones.

4.7. Isoformas especificas de tejido

A partir de los datos del trabajo de Pandey (del estudio ampliado) hemos podido analizar el *splicing* alternativo a nivel de tejido. Tan solo hemos sido capaces de detectar con claridad 14 genes que expresen isoformas especificas de tejido. A pesar de haber encontrado tan pocos casos, es interesante resaltar que de estas isoformas, 9 se expresan en corazón y 5 en cerebro. De las expresadas en corazón, 7 se generan a partir de eventos de exones homólogos y muchas son proteínas contráctiles del sarcómero.

4.8. Caracterización de isoformas principales

Como parte del estudio ampliado hemos analizado ocho experimentos proteómicos a gran escala y hemos comprobado que la inmensa mayoría de los genes expresan una isoforma principal a nivel de proteína, al margen del tipo de tejido o de linear celular.

Las isoformas principales detectadas a nivel de proteína coinciden con las predicciones de dos métodos ortogonales: APPRIS, basado en la conservación de proteína y CCDS, basado en la evidencia a nivel genómico. En el caso de aquellos genes para los que somos capaces de estimar la isoforma con los tres métodos, el acuerdo es de más del 99%.

El hecho de que la proteína de la isoforma principal coincida con la predicha por APPRIS valida la hipótesis de que la isoforma que se expresa a nivel de proteína sea aquella que tiene su estructura y función biológica más conservadas. Esto además implica que las predicciones de APPRIS podrían ser utilizadas para predecir las isoformas principales de los genomas de especies diferentes, lo que facilitaría el análisis de datos a gran escala. Ahora mismo ya se han incorporado a la base de datos de APPRIS 7 vertebrados, incluyendo ratón, rata, lince ibérico, zebrafish, cerdo, chimpancé y perro, y dos invertebrados, *Drosophila* y *C. Elegans.* Se están automatizando los procesos para poder incluir las predicciones de todos los genomas disponibles.

Un análisis reciente realizado en base a los datos generados para el proyecto de los 1,000 genomas (Clarke et al., 2012) remarca la importancia de la isoforma principal de APPRIS. Lin y Liu (Liu & Lin, 2015) demostraron en un trabajo reciente que los exones del conjunto de las isoformas principales de APPRIS tienen proporcionalmente menos mutaciones severas que el conjunto de exones que aparece en las isoformas alternativas. Esto avala la idea de que las isoformas principales de APPRIS (y por tanto de los experimentos de proteómica) son las isoformas más importantes de la célula y que la mayoría de isoformas alternativas, en el caso de ser traducidas a proteínas, tendrían un papel secundario.

4.9. Mejoras futuras en la sensibilidad y validación de péptidos

Con el objetivo de mejorar la sensibilidad estamos desarrollando un nuevo método que evita el problema del cálculo de la razón de falsos descubrimientos (FDR) cuando se utilizan múltiples experimentos de datos proteómicos a gran escala. Cuando se concatenan varios experimentos, los péptidos correctamente identificados emergen en varios experimentos. Sin embargo, los péptidos señuelo tienden a distribuirse de manera aleatoria (Reiter et al., 2009). El efecto de unir muchos experimentos proteómicos implica un aumento considerable de la FDR de proteína, cuando se calcula de la manera clásica. El algoritmo que estamos desarrollando calcula fehacientemente el FDR a nivel de gen o proteína sin verse afectado por el número de experimentos proteómicos que se estén utilizando en un mismo análisis. Creemos que la aplicación de este método supondrá también una mejora sustancial en la sensibilidad, a la hora de detectar la expresión de isoformas alternativas.

Por otro lado, en un trabajo reciente establecimos una serie de requisitos que entendemos necesarios a la hora de publicar resultados de proteogenómica (Ezkurdia *et al.*, 2015a). Entre ellos, el de la confirmación manual de las identificaciones automáticas, y la consideración de factores como la homología de péptidos y las modificaciones post-traduccionales que pudieran estar alterando los resultados. Estas medidas son de especial relevancia cuando se trata de validar péptidos asociados a proteínas que no han sido detectadas anteriormente (*missing proteins*). En ninguno de los dos análisis presentados en esta tesis se hizo una verificación manual de los espectros identificados, en el estudio piloto por falta de medios y en el estudio ampliado por no disponer en muchos casos de los ficheros *raw*. Sin embargo se utilizaron flujos de validación conservadores para paliar esta carencia. Creemos que es importante resaltar la importancia de seguir unos estándares de publicación y validación de datos proteómicos de cara a publicaciones futuras cómo los que se sugieren en dos revisiones recientes en el campo (Nesvizhskii, 2014; Omenn et al., 2015).

4.10. Cuantificación de isoformas alternativas

La diversidad de conjuntos de péptidos únicos y solapantes entre distintas isoformas alternativas hace que la cuantificación de las mismas sea especialmente compleja. En esta memoria debido a las limitaciones impuestas por el tipo de datos proteómicos analizados, solo hemos podido realizar una cuantificación aproximada de las isoformas a partir del recuento simple de PSMs. Poder estimar la abundancia de las diferentes isoformas en distintos tejidos y situaciones es de gran interés para entender mejor la función celular. El conjunto de genes que expresan más de una proteína que hemos detectado en los trabajos presentados podría ser una referencia útil de cara a desarrollar nuevos algoritmos de cuantificación y estrategias experimentales.

4.11. Discrepancias entre *splicing* de RNAs y *splicing* de proteínas

Los experimentos de *splicing* alternativo a nivel de transcrito, microarray y RNAseq demuestran la existencia de miles de variantes de *splicing*. Sin embargo, la evidencia de isoformas alternativas a nivel de proteína es más baja de la esperada. Los experimentos proteómicos identifican unas pocas isoformas alternativas que tienden a estar formadas a partir de la sustitución de exones homólogos y otros eventos, que no implican la fractura o pérdida del dominio funcional.

Como hemos visto en nuestros análisis el acuerdo entre la isoforma principal, las variantes únicas de CCDS extraídas a partir de información genómica y las predicciones de las isoformas principales de APPRIS basadas en conservación, demuestran que la gran mayoría de genes tiene una única isoforma dominante. Cuando comparamos el conjunto de isoformas principales detectadas a partir de proteómica con aquellos transcritos dominantes sugeridos por los datos de RNAseq, el acuerdo baja a un 82% en el mejor de los casos.

En base a lo observado en estos análisis parece claro que hay una gran diferencia entre lo que se puede detectar al nivel de transcrito y lo que se detecta al nivel de proteína.

Creemos que hay varios factores implicados en la discrepancia entre la expresión de transcritos y la expresión a nivel de proteína. Uno de los factores es la dificultad de detección de péptidos consecuencia de las limitaciones de la espectrometría de masas. Otra de las causas podría estar relacionada con los errores de los algoritmos de ensamblado de los transcritos de RNAseq (Hayer *et al.*, 2015). En este sentido, el conjunto de 3.000 genes a los que podemos asignar una isoforma principal a nivel de proteína podría servir de referencia para validar algoritmos de ensamblado de secuencias de RNAseq. Recientemente (Melé et al., 2015), se ha propuesto, basándose en la evidencia de RNAseq en tejidos normales, la posibilidad de que al menos una parte de los transcritos alternativos detectados por RNAseq pudiera ser simplemente ruido biológico. También es posible que la función de muchos transcritos no tenga nada que ver con la producción de proteínas (Lareau & Brenner, 2015).

Al margen de las diferencias que puedan existir entre la expresión de transcrito y la expresión a nivel de proteína, y al margen de los problemas de detección consecuencia de las limitaciones de la espectrometría de masas, creemos que puede haber otros factores implicados en esta discrepancia. Recientemente (Melé et al., 2015), se ha propuesto, basándose en la evidencia de RNAseq en tejidos normales, la posibilidad de que al menos una parte de los transcritos alternativos detectados por RNAseq pudiera ser simplemente ruido biológico. Otra de las causas podría estar relacionada con los errores de los algoritmos de ensamblado. En este sentido, el conjunto de 3.000 genes a los que podemos asignar una isoforma principal a nivel de proteína podría servir de referencia para validar algoritmos de ensamblado de secuencias de RNAseq.

Las mejoras en la tecnología de espectrometría de masas y de algoritmos de validación, junto con la aportación de nuevos experimentos proteómicos (como los que tienen en cuenta la diversidad de tejidos y líneas celulares), ayudarán en el esclarecimiento de esta discrepancia, esencial para entender mejor la maquinaria celular.

5. Conclusiones

1. La proteómica comparativa es útil para mejorar las anotaciones genómicas.
2. El número de isoformas alternativas detectadas a nivel de proteína es menor del esperado teóricamente.
3. La conservación funcional y estructural de dominios tiene un papel crucial en la evolución de las isoformas de *splicing* alternativo.
4. La sustitución de exones homólogos es el mecanismo más proclive a la generación de isoformas alternativas a nivel de proteínas.
5. El análisis riguroso de experimentos de proteómica a gran escala demuestra la existencia de una isoforma principal para la inmensa mayoría de los genes, permitiendo determinar cuál es dicha isoforma.
6. El elevado nivel de acuerdo de las isoformas principales, establecidas a partir de los resultados de proteómica con aquellas definidas por CCDS y APPRIS muestra que la isoforma proteómica principal es la más conservada y la que tiene más evidencia genómica.
7. La expresión de transcritos de isoformas alternativas difiere de forma notable de su expresión a nivel de proteína.
8. La traducción de isoformas alternativas parece sujeta a un mecanismo de regulación, capaz de explicar la diferente presencia de estas isoformas a nivel de transcrito y de proteína.
9. Para un 95% de los genes, los transcritos anotados como alternativos, de estarse expresando a nivel de proteína, lo estarían haciendo en bajas cantidades, en tejidos específicos o tendrían una vida media corta.

6. Materiales y métodos

6.1. Bases de datos de anotaciones genómicas

En esta memoria se utilizaron las bases de datos de anotaciones genómicas de las versiones 3C y 7 de GENCODE (humano) y la versión NCBIM37 de Ensembl (ratón) para los análisis del estudio piloto. Para el estudio ampliado se utilizó GENCODE 20 y GENCODE mouse M2 (ratón).

En la Figura 20, podemos ver el número de genes y transcritos anotados en las diferentes versiones de las bases de datos de anotaciones genómicas utilizadas en este trabajo. La versión 3C de GENCODE contenía 22550 genes codificantes que daban lugar a un total de 56217 transcritos codificantes de proteína. De estos genes, 10059 expresaban una única proteína y 12491 codificaban para varias isoformas a partir de *splicing* alternativo. La versión 7 de GENCODE tenía 60495 transcritos generados por 20687 genes codificantes, de los que 13346 expresaban más de una isoforma. En la versión 20 de GENCODE teníamos 92,341 transcritos que pertenecían a 19,906 genes codificantes, de los que 15,548 expresaban más de una isoforma.

La versión NCBIM37 de Ensembl para ratón era menos completa y tenía anotados un total de 22380 genes codificantes de los cuales, 9953 codificaban para un solo tránscrito y 12427 lo hacían para 43,815 tránscritos. La versión M2 tenía 22572 genes codificantes y 47394 transcritos codificantes.

La mayor parte de los análisis llevados acabo en el estudio piloto se hicieron sobre la versión 3C de GENCODE para humano y la versión NCBIM37 para ratón.

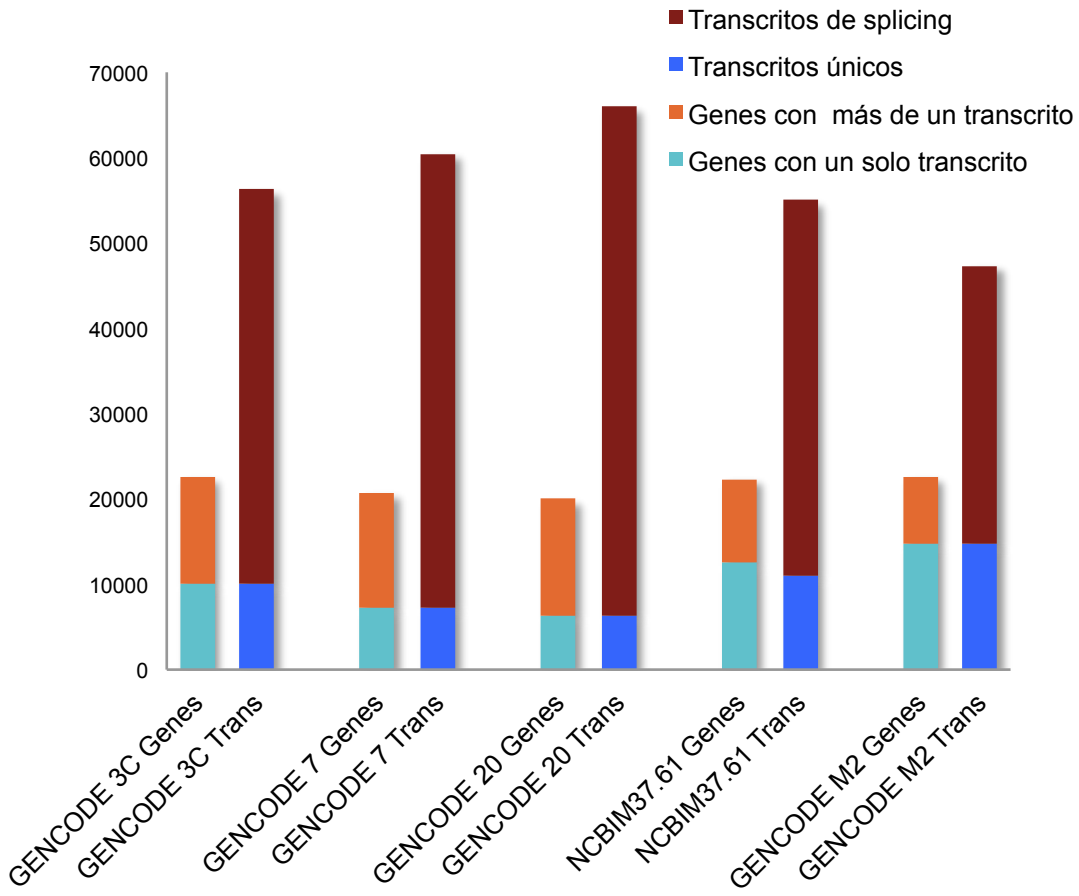


Figura 20. Fuentes de anotaciones genómicas utilizadas para el análisis de proteogenómica. Bases datos de anotaciones para humano GENCODE 3C, GENCODE 7 y GENCODE 20, y para ratón NCBIM37.61 y GENCODE M2. En las columnas podemos ver para las diferentes versiones utilizadas, el número de genes codificantes anotados con un sólo transcrito (azul claro), genes anotados con más de un tránscrito (naranja claro), tránscritos únicos (azul oscuro) y tránscritos generados a partir de splicing alternativo (marrón).

6.2. Base de datos de quimeras

El flujo de trabajo de proteogenómica se aplicó para buscar proteínas quiméricas. Para ello se utilizaron 5397 ARNm quiméricos procedentes de la base de datos ChimerDB (Kim *et al.*, 2010). Los transcritos se tradujeron a partir de 6 posibles marcos abiertos de lectura que dieron lugar a 32,382 proteínas quiméricas posibles. Para la búsqueda en los experimentos de espectrometría de masas solo se consideraron los péptidos que se encuentran en la zona de unión de los dos genes artífices de la proteína quimérica, siempre que hubiera al menos 3 aminoácidos en cada uno de los genes.

6.2. Repositorios de datos proteómicos

6.2.1. Repositorios de espectros para el primer estudio

El análisis del primer estudio se realizo a partir de las búsquedas de péptidos en los espectros de dos repositorios públicos. El repositorio de GPM (The Global Proteome Machine Organization) (Craig *et al.*, 2004) en el que se reanalizaron 5,809 ficheros de espectros en formato mzXML y el repositorio de PeptideAtlas (Desiere *et al.*, 2006) en el que se reanalizaron 52,019 ficheros de espectros en formato mzXML. Los ficheros se descargaron a través del sistema de descarga distribuido Tranche (tranche.proteomecommons.org) y del servidor de FTP ftp://ftp.thegpm.org/data/msms/. Al conjunto de péptidos obtenidos a partir de estas búsquedas y de su validación le hemos llamado *CNIO_proteodata* para referirnos a él en esta memoria.

Para el análisis de ratón en el primer estudio se utilizaron los espectros de los repositorios de GPM y PeptideAtlas. Los espectros utilizados para el análisis de ratón se obtuvieron a partir de 3509 ficheros mzXML depositados en PeptideAtlas y 2672 ficheros mzXML depositados en GPM.

El analisis de mosca se realizó a partir de los peptidos detectados en dos experimentos de proteómica a gran escala. En el primero se detectaron 32,729 peptidos y se identificaron

6,980 genes (Brunner *et al.*, 2007). En el segundo (Bodenmiller *et al.*, 2007) se encontraron 10,118 peptidos fosforilados que mapeaban a 3,472 genes. Los peptidos se mapearon a la base de datos de FlyBase (Tweedie *et al.*, 2009). Los detalles de los métodos empleados en el análisis se pueden encontrar en el estudio original (Tress *et al.*, 2008).

6.2.2. Repositorios de péptidos para el estudio ampliado

Para el análisis del estudio ampliado no se realizaron nuevas búsquedas, se utilizaron los péptidos encontrados en ocho fuentes de datos proteómicos. De estas, seis corresponden a los datos publicados en estudios de proteómica a gran escala pertenecientes a los análisis de Ezkurdia (Ezkurdia *et al.*, 2014a), Muñoz (Munoz *et al.*, 2011), Nagaraj (Nagaraj *et al.*, 2011), Geiger (Geiger *et al.*, 2012), Wilhelm (Wilhelm *et al.*, 2014) y Kim (Kim *et al.*, 2014) y las otras dos corresponden a los repositorios públicos de PeptideAtlas (Farrah *et al.*, 2012) y NIST ((http://peptide.nist.gov/). Al conjunto de péptidos filtrados de estas fuentes de datos le hemos llamado *Eight_proteodata* para referirnos a él en esta memoria.

Para el análisis de ratón en el estudio ampliado se utilizaron péptidos de NIST y de nuevos experimentos de PeptideAtlas además de péptidos nuevos obtenidos a partir de búsquedas de XTandem contra experimentos de PeptideAtlas y GPM.

6.3. Reanálisis de espectros de masas para el primer estudio

Las asignaciones espectro-péptido (PSM, del inglés Peptide-Spectrum Match) se obtuvieron a partir de la búsqueda de los peptidos de GENCODE 3C en los espectros de los repositorios de PeptideAtlas y GPM utilizando X!Tandem (Craig & Beavis, 2004).

Las búsquedas con X!Tandem se llevaron a cabo utilizando un cluster de Linux basado en un sistema de colas MPI (Massive Passing Interface). Para la mayoría de experimentos no se encontraron las especificaciones técnicas empleadas en la espectrometría de masas y se optó por utilizar parámetros de búsqueda estándar para el X!Tandem (tryptic cleavage specificity;

Mass tolerance of +- 20 ppm for precursor ions; +-0.4 Da for fragment ions; one missed cleavages permitted).

Las búsquedas se hicieron utilizando la base de datos de anotaciones GENCODE 3C, la base de datos reversa señuelo y cRAP (base de datos de contaminantes) de manera concatenada.

Se excluyeron todos los PSM con e-values por debajo de 0.01 para descartar péptidos falsos *a priori*, obteniendo un total de 950,684 PSM. Se descartaron todos los peptidos no trípticos que no fueran los C-terminales reduciéndose el conjunto a 918,484 péptidos trípticos.

Para medir el rendimiento se estimo la tasa de falsa detección (FDR) utilizado la estrategia de la base de datos señuelo (target-decoy) (Moore *et al.*, 2002). La base de datos señuelo se generó para los 62,943 transcritos anotados en GENCODE 3C. Las secuencias fueron reemplazadas con secuencias aleatorias del mismo tamaño y con los mismos porcentajes de composición de aminoácidos (www.matrixscience.com/downloads/decoy.pl.gz). Esta aproximación para la generación de la base de datos señuelo es una aproximación conservadora ya que también se generaron péptidos diferentes a partir de las secuencias que aparecían repetidas en GENCODE 3C. La base de datos dió lugar a 1,7 millones de péptidos trípticos, mientras que su base equivalente aleatoria generó 4 millones de péptidos diferentes.

6.3.1. Flujo de trabajo de proteogenómica

En la Figura 21 se muestra el flujo de trabajo para el análisis de proteogenómica:

- Se utilizaron las bases de datos de anotaciones genómicas de Gencode 3c y 7 (humano), NCBIM37 (ratón) y ChimerDB (quimeras humanas) en formato fasta. Se utilizó CD-HIT para eliminar secuencias repetidas, se añadió una lista de contaminantes comunes y se generaron las bases de datos señuelo para realizar una búsqueda concatenada.

- Se crearon los repositorios de datos de proteómica y se convirtieron los ficheros de espectros a formato mzXML desde *raw* cuando fue necesario. Las búsquedas de péptidos se realizan utilizando x!Tandem en un cluster con un sistema de colas MPI.
- Se extrajeron los resultados de las búsquedas para su posterior análisis. Los péptidos se validaron y ordenaron en función de su FDR para su posterior análisis.
- A parir del umbral escogido, se mapean los péptidos en función de las base de datos de anotaciones genómicas.

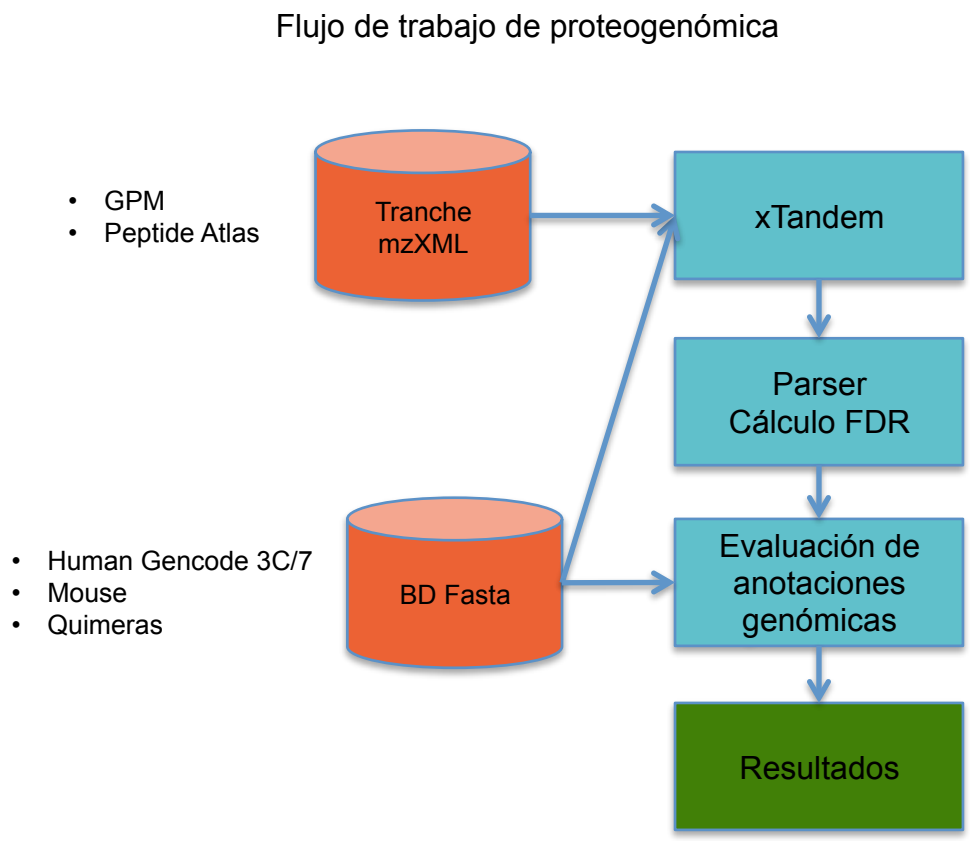


Figura 21. Flujo de trabajo del análisis para el estudio piloto. Las búsquedas en los espectros de los repositorios públicos de GPM y PeptideAtlas se llevaron a cabo a partir de las bases de datos de anotaciones genómicas de GENCODE y Ensembl en humano y ratón utilizando el buscador xTandem. Los resultados se parsearon y procedió al cálculo de la tasa de falso descubrimiento (FDR). Después a partir de varios programas se procedió al mapeo y análisis de péptidos para la obtención de resultados.

6.4. Validación de péptidos identificados para el estudio ampliado

La validación de los péptidos provenientes del estudio ampliado se hizo en base a la aplicación de una serie de filtros para garantizar el menor número posible de falsos positivos.

Para la validación de los péptidos resultantes de los experimentos de Geiger (Geiger *et al.*, 2012) y Nagaraj (Nagaraj *et al.*, 2011) (con un FDR reportado por sus autores del 1%) se aplicaron las siguientes reglas: que estuvieran identificados en al menos dos conjuntos de datos y que tuvieran una puntuación de Andromeda mayor que 100. La razón para aplicar este filtro fue asegurarse que con una puntuación de Andromeda mayor que 100, estos péptidos habrían sido probablemente identificados también por Mascot (Koenig *et al.*, 2008), de este modo simulamos que los péptidos habrían sido detectados por dos motores de búsqueda (acorde con la exigencia impuesta en el resto de filtros) lo que en teoría aumentaría la fiabilidad (Shteynberg *et al.*, 2013).

Para la validación los péptidos provenientes del experimento de Kim (Kim *et al.*, 2014) (con un FDR de PSM reportado por sus autores del 1%) se tuvieron en cuenta únicamente los péptidos que habían sido detectados por ambos motores de búsqueda utilizados, Sequest y Mascot. El propósito de este filtro era reducir el error que suele cometerse cuando se emplean los resultados de la unión de dos motores de búsqueda (Shteynberg *et al.*, 2013).

Para la validación los resultados del análisis de Wilhelm (Wilhelm *et al.*, 2014) (con un FDR a nivel de PSM reportado por sus autores del 1% y un FDR a nivel de péptido del 5%) se tuvieron en cuenta únicamente los péptidos que habían sido detectados por ambos motores de búsqueda utilizados, Andromeda y Mascot. El proposito de este filtro era reducir el error que suele cometerse cuando se emplean los resultados de la unión de dos motores de búsqueda. Dado que el número de falsos positivos seguía siendo elevado se aplicó la misma regla que en el caso de Geiger y Nagaraj, teniendo en cuenta solo aquellos con puntuaciones de Andrómeda mayores que 100.

Para la validación de los péptidos del repositorio de NIST se filtraron todos aquellos péptidos detectados con un solo motor de búsqueda (NIST emplea cinco motores de búsqueda: Sequest, Andromeda, Mascot, X!Tandem y OMMSA).

Para los péptidos del primer estudio (0.1% de FDR a nivel de péptido), los péptidos de Muñoz (Munoz *et al.*, 2011) (1% de FDR a nivel de péptidos) y los de PeptideAtlas (0.0002% de FDR a nivel de PSM) no se aplico ningún filtro adicional.

Todos los péptidos fueron sometidos a los siguientes filtros: se descartaron todos los péptidos no trípticos y semi-tripticos, se incluyeron solo los péptidos con digestiones parciales (*miss-cleavage*) en los casos en los que existía un un subpéptido sin digestión parcial, (para los péptidos de Wilhelm y Nagaraj se aplicaron los filtros equivalentes teniendo en cuenta las enzimas de digestión LysC y la quimotripsina y las enzimas GluC o LysC respectivamente; para los péptidos del primer estudio, PeptideAtlas y NIST para los que no se sabía la enzima empleada se asumió que habían sido digeridos con tripsina), los residuos de leucina e isoleucina se trataron indistintamente a la hora de mapearlos contra la GENCODE, se descartaron todos aquellos péptidos que mapeaban a más de un gen.

6.5. Identificación de isoformas alternativas

En el primer estudio, los péptidos validados a partir del flujo de proteogenómica se mapearon a las proteínas en GENCODE 3C. Se seleccionaron solo aquellos péptidos capaces de distinguir de manera inequívoca una isoforma o un subconjunto de isoformas del mismo gen. A partir de los datos encontrados se generó una base de datos con la información relativa a cada uno de los péptidos detectados y sus correspondientes mapeos en la base de datos de anotaciones genómicas.

En el estudio ampliado solo se tuvieron en cuenta los péptidos que habían sido identificado en dos o más de los ocho conjuntos de datos de *Eight_proteodata*. Los péptidos se mapearon en todas las isoformas de GENCODE 20 (excluyendo los genes *read-through* y los pseudoautosómicos), 92,341 transcritos pertencientes a 19,906 genes codificantes.

En el caso de ratón los péptidos se mapearon en 51,610 transcritos de GENCODE mouse M2 (equivalente a Ensembl74) pertenecientes a 22,645 genes codificantes. En el caso de ratón solo había 10,607 genes con más de una isoforma anotada.

En la tabla 6 se han incluido todos los genes con más de una isoforma detectados en el estudio ampliado.

Gen	Principal	Alternativa	Alt peps	Tipo	Efecto Pfam	Ratón
ACOX1	ENST00000301608	ENST00000293217	2	HomolEx	No effect	Mouse
ACP1	ENST00000272065	ENST00000272067 ...	6	HomolEx	No effect	
ACTN1	ENST00000394419	ENST00000193403	7	HomolEx	No effect	
ACTN4	ENST00000252699	ENST00000440400	3	HomolEx	No effect	Mouse
ADAR	ENST00000368474	ENST00000529168	4	Indel	No effect	
ADD1	ENST00000264758	ENST00000398123	2	C-term	No effect	
ANK2	ENST00000357077	ENST00000264366 ...	4	**Indel**	No effect	Mouse
ANK3	ENST00000373827	ENST00000280772 ...	2	**Indel**	No effect	
ANXA6	ENST00000354546	ENST00000521512	12	Insert	No effect	Mouse
AP1B1	ENST00000405198 ...	ENST00000432560 ...	5	NAGNAG	No effect	
AP2A2	ENST00000332231	ENST00000448903	4	NAGNAG	No effect	
API5	ENST00000378852	ENST00000531273 ...	3	C-term	No effect	
ARFGAP3	ENST00000263245	ENST00000453516	6	**Indel**	No effect	
ARFIP1	ENST00000353617 ...	ENST00000618090 ...	8	**Indel**	No effect	
ARHGDIA	ENST00000269321 ...	ENST00000584461	2	C-term	Broken	
ARMC10	ENST00000323716 ...	ENST00000441711 ...	6	**Indel**	No effect	
ASPH	ENST00000379454	ENST00000445642	2	C-term	Broken	
ATE1	ENST00000369040 ...	ENST00000224652 ...	2	HomolEx	No effect	
ATF7IP	ENST00000261168	ENST00000536444	2	NAGNAG	No effect	
ATL1	ENST00000358385	ENST00000441560	2	**Indel**	No effect	
ATP2B1	ENST00000428670 ...	ENST00000359142	2	HomolEx	Broken	
ATP2B4	ENST00000357681	ENST00000367218 ...	8	HomolEx	Broken	
BRAF	ENST00000288602	ENST00000496384	2	HomolEx	No effect	
BTF3L4	ENST00000313334	ENST00000489308	2	C-term	Broken	
C17orf49	ENST00000439424 ...	ENST00000552402	4	Indel	No effect	
C18orf25	ENST00000619301	ENST00000615052	3	**Indel**	No effect	
CA6	ENST00000377443	ENST00000377436	2	C-term	No effect	
CALD1	ENST00000361675	ENST00000422748	2	**Indel**	Broken	
CALU	ENST00000249364 ...	ENST00000449187 ...	9	HomolEx	No effect	Mouse
CAMSAP3	ENST00000160298	ENST00000615537 ...	2	Indel	No effect	
CAPZB	ENST00000264202 ...	ENST00000375142	9	HomolEx	No effect	
CASK	ENST00000421587	ENST00000442742 ...	3	**Indel**	No effect	
CASP8	ENST00000432109 ...	ENST00000392258 ...	2	C-term	Lost	
CAST	ENST00000508830	ENST00000511049 ...	2	NAGNAG	No effect	
CBFB	ENST00000412916	ENST00000290858 ...	5	C-term	No effect	
CD46	ENST00000354848	ENST00000367042	8	C-term	No effect	
CDC42	ENST00000400259 ...	ENST00000315554	7	HomolEx	No effect	Mouse
CDK13	ENST00000340829	ENST00000181839	2	indel	No effect	
CDKN2A	ENST00000498124 ...	ENST00000361570 ...	7	Two proteins	Lost	
CHTF8	ENST00000306585	ENST00000522497 ...	5	Two proteins	Lost	
CLASP1	ENST00000455322 ...	ENST00000263710 ...	4	**Indel**	No effect	
CLINT1	ENST00000523908	ENST00000411809 ...	4	**Indel**	No effect	
CLIP2	ENST00000223398 ...	ENST00000361545	2	Indel	No effect	
CLTB	ENST00000310418	ENST00000345807	2	Indel	Broken	
CNBP	ENST00000422453 ...	ENST00000451728	7	NAGNAG	No effect	Mouse
CNOT1	ENST00000317147	ENST00000569240 ...	3	**Indel**	No effect	
COL5A1	ENST00000618395	ENST00000371817	2	HomolEx	No effect	

COL6A2	ENST00000300527	ENST00000397763	5	C-term	No effect	
COL6A3	ENST00000295550	ENST00000472056 ...	5	**Indel**	lost	
CP	ENST00000264613	ENST00000455472	2	Indel	No effect	
CRMP1	ENST00000397890	ENST00000324989	2	N-term	No effect	
CTB-50L17.10	ENST00000616600	ENST00000621835	3	NAGNAG	No effect	
CTBP1	ENST00000290921 ...	ENST00000503594	2	NAGNAG	No effect	
CUX1	ENST00000292535 ...	ENST00000292538 ...	23	C-term	Swap	Mouse
DBNL	ENST00000448521	ENST00000494774 ...	8	NAGNAG	No effect	
DCTN1	ENST00000409868 ...	ENST00000361874 ...	3	**Indel**	No effect	
DDX39B	ENST00000396172 ...	ENST00000428450	2	**Indel**	Broken	
DDX46	ENST00000354283	ENST00000452510	3	NAGNAG	No effect	
DFFA	ENST00000377038	ENST00000377036	6	C-term	No effect	
DLGAP5	ENST00000247191	ENST00000395425	2	C-term	No effect	
DMD	ENST00000378723	ENST00000357033	5	N-term	Broken	
DMTN	ENST00000265800 ...	ENST00000381470 ...	2	Indel	No effect	
DNAJC5	ENST00000360864	ENST00000470551	2	C-term	No effect	
DNM1	ENST00000372923 ...	ENST00000475805 ...	5	HomolEx	No effect	Mouse
DNM2	ENST00000355667	ENST00000389253 ...	5	HomolEx	No effect	
DST	ENST00000361203	ENST00000370765	4	C-term	Swap	
DUT	ENST00000331200	ENST00000455976 ...	6	N-term	No effect	Mouse
DYNC1I2	ENST00000409317	ENST00000340296 ...	2	**Indel**	No effect	
EEF1D	ENST00000532741 ...	ENST00000524624 ...	8	Indel	No effect	Mouse
EIF2B4	ENST00000347454	ENST00000445933	7	NAGNAG	No effect	
EIF4G1	ENST00000342981	ENST00000346169	5	NAGNAG	No effect	
EPB41L1	ENST00000338074	ENST00000202028 ...	7	**Indel**	No effect;	
EPB41L3	ENST00000540638	ENST00000342933 ...	10	Indel	No effect	
ESYT1	ENST00000267113	ENST00000394048	2	**Indel**	Broken	
FBLN1	ENST00000327858	ENST00000262722 ...	18	HomolEx	No effect	Mouse
FLAD1	ENST00000315144 ...	ENST00000368431	2	C-term	Lost	
FLII	ENST00000327031	ENST00000545457 ...	7	NAGNAG	No effect	
FLNA	ENST00000422373 ...	ENST00000369850 ...	4	**Indel**	Broken	
FMN1	ENST00000334528	ENST00000559047	2	N-term	No effect	
FMNL3	ENST00000550488	ENST00000335154 ...	2	HomolEx	No effect	
FMR1	ENST00000370475	ENST00000439526	5	indel	No effect	
FN1	ENST00000359671	ENST00000336916 ...	4	Indel	**Lost**	
FOXP1	ENST00000475937 ...	ENST00000493089	3	NAGNAG	No effect	
FUBP1	ENST00000370768	ENST00000294623 ...	6	NAGNAG	No effect	
FYN	ENST00000538466 ...	ENST00000368667 ...	6	HomolEx	No effect	
G3BP2	ENST00000359707 ...	ENST00000357854	3	Indel	No effect	
G6PD	ENST00000393562 ...	ENST00000439227 ...	4	NAGNAG	No effect	
GANAB	ENST00000346178	ENST00000356638 ...	5	**Indel**	No effect	
GLS	ENST00000338435	ENST00000320717 ...	20	C-term	Lost	Mouse
GNAO1	ENST00000262493	ENST00000262494	6	HomolEx	No effect	
GSN	ENST00000373818	ENST00000436847 ...	4	N-term	No effect	
H2AFY	ENST00000511689 ...	ENST00000312469	5	HomolEx	No effect	Mouse
HBS1L	ENST00000367837	ENST00000367822	8	C-term	Lost	
HCFC1	ENST00000310441	ENST00000369984	6	**Indel**	No effect	
HM13	ENST00000340852	ENST00000398174 ...	3	**Indel**	No effect	

HMGA1	ENST00000311487 ...	ENST00000374116 ...	2	**Indel**	No effect	
HNRNPC	ENST00000556897 ...	ENST00000554455 ...	4	Indel	No effect	
HNRNPD	ENST00000353341	ENST00000352301 ...	4	**Indel**	No effect	
HNRNPH3	ENST00000265866	ENST00000354695	2	**Indel**	No effect	
HNRNPK	ENST00000376263 ...	ENST00000457156	8	**Indel**	No effect	
HNRNPR	ENST00000374612 ...	ENST00000374616 ...	11	NAGNAG	No effect	
HNRNPU	ENST00000444376	ENST00000283179	6	**Indel**	No effect	
IKBIP	ENST00000299157	ENST00000342502	43	HomolEx	No effect	Mouse
ILF3	ENST00000449870 ...	ENST00000590261 ...	7	NAGNAG	No effect	
IMMT	ENST00000449247	ENST00000410111 ...	9	Indel	no effect	
IMPDH2	ENST00000326739	ENST00000429182	2	**Indel**	No effect	
ITGA6	ENST00000264107	ENST00000442250 ...	3	HomolEx	No effect	
ITGA7	ENST00000257879	ENST00000553804 ...	3	HomolEx	No effect	
ITIH4	ENST00000266041 ...	ENST00000406595 ...	4	**Indel**	No effect	
ITPR1	ENST00000443694 ...	0	2	indel	Broken	
KHK	ENST00000260598	ENST00000260599 ...	4	HomolEx	No effect	
KIAA1468	ENST00000256858	ENST00000398130	5	HomolEx	No effect	
KIF1B	ENST00000377081 ...	ENST00000377083 ...	4	C-term	Lost	
KIF23	ENST00000260363 ...	ENST00000352331 ...	3	Subst	No effect	
KLC1	ENST00000555836 ...	ENST00000557450	4	C-term	No effect	
KNG1	ENST00000265023	ENST00000287611 ...	12	C-term	No effect	
KTN1	ENST00000459737 ...	ENST00000438792 ...	2	indel	no effect	
LAMP2	ENST00000200639	ENST000003713355	2	HomolEx	No effect	
LDB3	ENST00000361373	ENST00000429277 ...	10	HomolEx	noeffect	
LMNA	ENST00000368300	ENST00000368301 ...	13	C-term	No effect	
LSM14B	ENST00000279068	ENST00000370915	3	C-term	Lost	
LSR	ENST00000605618 ...	ENST00000602122	3	NAGNAG	No effect	
MAP4	ENST00000426837	ENST00000429422	3	C-term	No effect	
MAP4K4	ENST00000350878 ...	ENST00000347699 ...	2	NAGNAG	No effect	
MAPK14	ENST00000229794	ENST00000229795	8	HomolEx	No effect	
MAPT	ENST00000262410	ENST00000351559	7	**Indel**	No effect	
MASP1	ENST00000296280	ENST00000337774	12	HomolEx	No effect	
MAT2B	ENST00000321757	ENST00000280969	6	N-term	No effect	
MATN2	ENST00000254898	ENST00000521689	2	Indel	No effect	
MAZ	ENST00000568282 ...	ENST00000568544 ...	2	C-term	No effect	
MBNL1	ENST00000282486 ...	ENST00000355460 ...	4	**Indel**	No effect	
MFF	ENST00000353339 ...	ENST00000354503 ...	2	**Indel**	Broken	
MFI2	ENST00000296350	ENST00000296351	2	C-term	Broken	
MKI67	ENST00000368654 ...	ENST00000368653	7	**Indel**	No effect	
MLLT4	ENST00000447894	ENST00000366806	4	NAGNAG	No effect	
MOCS2	ENST00000396954	ENST00000450852 ...	12	Two proteins (Swap)	Swap	
MPRIP	ENST00000341712 ...	ENST00000313485	10	**Indel**	No effect	
MPZL1	ENST00000359523	ENST00000367853 ...	3	C-term	No effect	
MRPL43	ENST00000370236 ...	ENST00000299179 ...	2	C-term	No effect	
MXRA7	ENST00000449428	ENST00000355797	5	C-term	No effect	
MYEF2	ENST00000324324	ENST00000267836 ...	3	**Indel**	No effect	
MYH11	ENST00000300036	ENST00000396324 ...	8	C-term	No effect	

MYL6	ENST00000548293	ENST00000547649	5	HomolEx	No effect	
MYO1B	ENST00000304164	ENST00000339514	4	**Indel**	No effect	
NAA50	ENST0000024092	ENST00000493900	2	NAGNAG	No effect	
NASP	ENST00000350030	ENST00000351223 ...	4	**Indel**	No effect	
NEBL	ENST00000377122 ...	ENST00000417816	30	N-term (Swap)	Swap	
NEDD4L	ENST00000382850	ENST00000400345 ...	3	**Indel**	No effect	
NOLC1	ENST00000605788 ...	ENST00000488254	4	NAGNAG	No effect	
NPM1	ENST00000296930 ...	ENST00000393820	2	C-term	No effect	
NSFL1C	ENST00000216879	ENST00000476071	2	NAGNAG	No effect	
NUDT4	ENST00000415493	ENST00000337179 ...	2	NAGNAG	No effect	
NUMA1	ENST00000358965	ENST00000393695	8	**Indel**	No effect	
OGT	ENST00000373719	ENST00000373701	6	**Indel**	No effect	
P4HA1	ENST00000263556	ENST00000394890	7	HomolEx	No effect	
P4HA2	ENST00000401867	ENST00000360568	4	HomolEx	No effect	
PALLD	ENST00000512127	ENST00000261509 ...	15	N-term	Lost	
PALM	ENST00000338448	ENST00000264560	4	**Indel**	Broken	
PCBP2	ENST00000359462	ENST00000439930 ...	4	NAGNAG	No effect	
PDLIM3	ENST00000284771	ENST00000284770	12	HomolEx	No effect	
PDLIM5	ENST00000317968	ENST00000514743 ...	15	Indel	no effect	
PES1	ENST00000402284 ...	ENST00000335214	3	**Indel**	No effect	
PFN2	ENST00000452853	ENST00000239940 ...	12	HomolEx	No effect	
PHF1	ENST00000495509 ...	ENST00000374516	3	C-term	lost	
PHF6	ENST00000370803 ...	ENST00000370799 ...	5	NAGNAG	No effect	
PKM	ENST00000319622 ...	ENST00000335181 ...	31	HomolEx	No effect	Mouse
PLCD1	ENST00000334661	ENST00000463876	5	N-term	No effect	
PLEC	ENST00000322810	ENST00000345136	15	**N-term**	No effect	Mouse
PMS1	ENST00000418224 ...	ENST00000452382	2	N-term	Broken	
PNKD	ENST00000273077	ENST00000248451	8	Two proteins	lost	
POLDIP3	ENST00000252115	ENST00000348657 ...	5	**Indel**	No effect	
POSTN	ENST00000379749	ENST00000379747 ...	6	**Indel**	No effect	
PPHLN1	ENST00000395580 ...	ENST00000395568 ...	10	C-term	No effect	
PPIP5K2	ENST00000321521 ...	ENST00000509597 ...	2	indel	No effect	
PPP1R12A	ENST00000261207 ...	ENST00000547330	4	**Indel**	No effect	
PPP1R12B	ENST00000608999	ENST00000391959	3	N-term	No effect	
PPP2R4	ENST00000337738	ENST00000358994 ...	2	**Indel**	Broken	
PPP3CA	ENST00000394854	ENST00000394853 ...	2	**Indel**	No effect	
PRKG1	ENST00000373980	ENST00000373985	6	HomolEx	No effect	
PSMD4	ENST00000368884	ENST00000368881	5	NAGNAG	No effect	Mouse
PTBP2	ENST00000426398 ...	ENST00000370197 ...	3	**Indel**	No effect	
PUF60	ENST00000349157	ENST00000527197 ...	5	Indel	No effect	
PUM1	ENST00000257075	ENST00000440538 ...	6	NAGNAG	No effect	
PVRL2	ENST00000252483	ENST00000252485	3	HomolEx	No effect	
RABGAP1L	ENST00000251507	ENST00000347255 ...	7	Two proteins	Lost	Mouse
RAP1GDS1	ENST00000339360 ...	ENST00000453712	4	NAGNAG	No effect	
RBM10	ENST00000329236	ENST00000345781	4	**Indel**	No effect	
RBM23	ENST00000399922	ENST00000359890	2	NAGNAG	No effect	
RBM26	ENST00000438724	ENST00000267229 ...	5	NAGNAG	No effect	

RBM7	ENST00000375490	ENST00000540163	2	NAGNAG	No effect	
RBPMS	ENST00000320203 ...	ENST00000339877 ...	2	C-term	No effect	
REPS1	ENST00000258062 ...	ENST00000367663 ...	3	**Indel**	No effect	
RNASE4/ANG	0	0	14	HomolEx	No effect	
RTN1	ENST00000267484	ENST00000342503	4	N-term	No effect	Mouse
RTN3	ENST00000377819 ...	ENST00000537981 ...	11	**Indel**	No effect	
RTN4	ENST00000337526	ENST00000357732 ...	6	Indel; indel	No effect	
RUFY3	ENST00000381006	ENST00000417478	3	N-term	No effect	
SBSN	ENST00000452271	ENST00000518157	4	**Indel**	No effect	
SCAMP3	ENST00000302631	ENST00000355379	2	Indel	No effect	
SDCBP	ENST00000413219 ...	ENST00000447182 ...	3	NAGNAG	No effect	
SEC16A	ENST00000313050	ENST00000290037 ...	2	**Indel**	No effect	
SERBP1	ENST00000370995	ENST00000361219 ...	8	**Indel**	No effect	
SERPINB13	ENST00000269489	ENST00000344731	2	NAGNAG	No effect	
SF1	ENST00000377394	ENST00000377387 ...	10	Cterm	No effect;	
SIRPB1	ENST00000381605	ENST00000568365 ...	2	HomolEx	No effect	
SLC25A3	ENST00000188376 ...	ENST00000228318 ...	4	HomolEx	No effect	
SLC4A4	ENST00000340595	ENST00000264485	2	N-term	No effect	
SLC9A3R2	ENST00000424542	ENST00000566198	2	N-term	Lost	
SMARCA4	ENST00000429416 ...	ENST00000413806 ...	2	Indel	No effect	
SNAP25	ENST00000254976	ENST00000304886	7	HomolEx	No effect	
SORBS1	ENST00000354106	ENST00000361941 ...	7	Indel	No effect	
SORBS2	ENST00000437304	ENST00000319471 ...	7	N-term; Subst	no effect	
SPAG9	ENST00000262013	ENST00000357122 ...	4	**Indel**	No effect;	
SPP1	ENST00000395080 ...	ENST00000237623	2	Indel	Broken	
SPTAN1	ENST00000372731	ENST00000372739 ...	10	Indel	No effect	Mouse
SPTBN1	ENST00000356805	ENST00000333896	8	N-term	No effect	Mouse
STRN3	ENST00000355683	ENST00000357479 ...	3	**Indel**	No effect	
STX16	ENST00000371141	ENST00000355957	3	Indel	No effect	
STXBP1	ENST00000373299	ENST00000373302	4	HomolEx	No effect	Mouse
SUGT1	ENST00000343788	ENST00000310528	6	**Indel**	No effect	
SYNCRIP	ENST00000369622	ENST00000355238 ...	3	C-term	No effect	
SYNJ1	ENST00000433931	ENST00000357345 ...	2	C-term	No effect	
SYNM	ENST00000336292	ENST00000594047 ...	3	**Indel**	No effect	
TAOK2	ENST00000279394	ENST00000308893 ...	5	C-term	No effect	
TIAL1	ENST00000369093	ENST00000436547 ...	5	**Indel**	No effect	
TMPO	ENST00000266732	ENST00000556029	65	C-term	Lost	Mouse
TNC	ENST00000350763	ENST00000341037 ...	2	**Indel**	Lost	
TOR1AIP1	ENST00000606911	ENST00000528443 ...	6	NAGNAG	No effect	
TPM1	ENST00000559397	ENST00000267996 ...	30	HomolEx	No effect	Mouse
TPM2	ENST00000378292	ENST00000329305 ...	25	HomolEx	No effect	Mouse
TPM3	ENST00000368533	ENST00000330188	6	HomolEx	No effect	Mouse
TPM4	ENST00000300933	ENST00000344824	35	HomolEx	No effect	
TRIM33	ENST00000450349 ...	ENST00000369543	4	**Indel**	No effect	
TSC22D1	ENST00000261489	ENST00000458659	2	N-term	No effect	
TSC22D3	ENST00000372384 ...	ENST00000372397 ...	2	N-term	No effect	
TSTD1	ENST00000423014	ENST00000368023	2	**Indel**	no effect	

TTN	ENST00000615779 ...	ENST00000360870	17	HomolEx	Lost	
UBQLN1	ENST00000376395	ENST00000257468	4	**Indel**	No effect	Mouse
UGT1A genes	0	0	21	HomolEx	No effect	Mouse
VAPA	ENST00000400000	ENST00000340541	4	**Indel**	No effect	
VDAC3	ENST00000022615	ENST00000521158 ...	4	NAGNAG	No effect	
YBX3	ENST00000228251	ENST00000279550	3	**Indel**	No effect	
ZC3HAV1	ENST00000242351	ENST00000464606	2	**Indel**	No effect	
ZMYND8	ENST00000262975 ...	ENST00000352431 ...	2	**Indel**	No effect	
ZNF185	ENST00000449285	ENST00000370268 ...	2	NAGNAG	No effect	
ZNF451	ENST00000370706 ...	ENST00000370708	2	C-term	Swap	

Tabla X. Genes detectados con más de una isoforma alternativa en proteómica.

En la tabla se muestran los genes de splicing alternativo detectados. En las columnas se muestran: el nombre del gen (Gen), el transcrito para el que se ha encontrado más evidencia de péptidos (Principal) (para los casos en los que no se pudo distinguir y hay más de uno, se indica con "…"), el transcrito alternativo detectado (Alternativo)) (en los casos en los que hay más de un posible transcrito solo se ha puesto uno y se indica con "…"), numero de péptidos que mapean al transcrito alternativo (Alt peps), tipo de evento de splicing identificado para el transcrito alternativo que aparece en la coluna (Tipo: exón homólogo *HomolEX,* deleción *Indel,* inserción *Insert,* NAGNAG, c terminal *C-term,* dos proteínas *Two proteins*), el efecto del evento sobre el dominio Pfam (efecto Pfam), casos en los que se encuentra el evento de splicing equivalente en los datos proteómicos de ratón.

6.6. Cálculo de niveles de expresión en el primer estudio

Los niveles de expresión a nivel de transcrito se calcularon a partir de la base de datos Huge Index database (Haverty *et al.*, 2002). Esta base de datos, contenía datos de 19 tipos diferentes de tejidos. Todas las muestras de tejidos utilizadas procedían de muestras de humanos sanos y los niveles de expresión fueron calculados bajo los mismos procesos de normalización y escalado, siendo los valores de los diferentes conjuntos de genes comparables entre si. La base de datos de Huge Index contenía la expresión de más de 7000 genes humanos.

6.7. Efecto de eventos de *splicing* en dominios Pfam para el studio ampliado.

Las búsquedas de dominios Pfam se llevaron a cabo utilizando el software Pfamscan (Bateman *et al.*, 2002).

Para el cálculo del efecto de eventos de splicing en dominios Pfam en el estudio ampliado se seleccionó la isoforma principal de APPRIS para cada gen, como referencia a la hora de determinar el evento de *splicing*. Se utilizó PfamScan para anotar los dominios en los trancritos y cuantificar los casos donde el dominio Pfam se rompía (perdida de cinco o más amino ácidos) o se perdía completamente.

6.8. Identificación de genes con exones homólogos para el estudio ampliado

La identificación automática de un conjunto de genes con exones homólogos mutuamente excluyentes se hizo a partir de la base de datos Ensembl (Silva *et al.*, 2013) versión 78. Para cada gen se seleccionó el transcrito referencia utilizando la secuencia más larga. Se seleccionaron aquellos exones mutuamente excluyentes de los transcritos de cada gen con respecto al transcrito referencia. Se escogieron los exones de mas de 30 pares de bases, se obtuvieron sus secuencias de amino ácidos y se compararon con BLAST v2.2.25 (Altschul *et al.*, 1990) utilizando un e-value de 0.005. La validación de los exones homólogos potenciales se hizo en función de si el exón ocupaba una posición equivalente con respecto al transito alternativo. Se incluyeron también casos adicionales identificados a partir de BLATP que no habían sido detectados con este método. A partir de una inspección visual se validaron 157 genes con exones homólogos mutuamente excluyentes.

7. Bibliografía

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp 1061–1073 England.

Abascal, F., Ezkurdia, I., Rodriguez-Rivas, J., Rodriguez, J. M., Del Pozo, A., Vazquez, J., Valencia, A. & Tress, M. L. (2015a). Alternatively Spliced Homologous Exons Have Ancient Origins and Are Highly Expressed at the Protein Level.(Käll, L., Ed) *PLoS computational biology*, 11(6), p e1004325 Public Library of Science.

Abascal, F., Tress, M. & Valencia, A. (2015b). The evolutionary fate of alternatively spliced homologous exons after gene duplication. *Genome Biology and Evolution*, 7(6), pp evv076–1403 Oxford University Press.

Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928), pp 198–207.

Ahrens, C. H., Brunner, E., Qeli, E., Basler, K., Aebersold, R., Ahrens, C. H., Brunner, E., Qeli, E., Basler, K. & Aebersold, R. (2010). Generating and navigating proteome maps using mass spectrometry. *Nature Reviews. Molecular Cell Biology*, 11(11), p 789 Nature Publishing Group.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp 403–410.

Baker, K. E. & Parker, R. (2004). Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Current opinion in cell biology*, 16(3), pp 293–299.

Barash, Y. E. A. & Barash, Y. E. A. (2010). Deciphering the splicing code. *Nature*, 465(7294), pp 53–59.

Barberan-Soler, S., Lambert, N. J. & Zahler, A. M. (2009). Global analysis of alternative splicing uncovers developmental regulation of nonsense-mediated decay in C. elegans. *RNA*, 15(9), pp 1652–1660.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. L. (2002). The Pfam Protein Families Database. *Nucleic Acids Research*, 30(1), pp 276–280 Oxford University Press.

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C. & Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal*, 33(9), pp 981–993 EMBO Press.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J. & Aebersold, R. (2011). The quantitative proteome of a human cell line. *Molecular systems biology*, 7(1) Nature Publishing Group.

BehmAnsmant, I., Kashima, I., Rehwinkel, J., Sauliere, J., Wittkopp, N. & Izaurralde, E. (2007). mRNA quality control: An ancient machinery recognizes and degrades mRNAs with nonsense codons. *FEBS Lett*, 581(15), pp 2845–2853.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucl. Acids Res*, 28(1), pp 235–242.

Black, D. L. (2000). Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology. *Cell*, 103(3), pp 367–370.

Black, D. L. (2003a). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev*

Biochem, 72, pp 291–336 United States.
Black, D. L. (2003b). Mechanisms of Alternative Pre-Messenger RNA Splicing. *dx.doi.org*, 72(1), pp 291–336 Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA.
Blencowe, B. J. & Graveley, B. R. (2008). *Alternative Splicing in the Postgenomic Era*. 1. ed Springer. ISBN 9780387773735.
Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D., Lam, H., Schmidt, A., Rinner, O., Mueller, L. N., Shannon, P. T., Pedrioli, P. G., Panse, C., Lee, H.-K., Schlapbach, R. & Aebersold, R. (2007). PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Molecular systems biology*, 3, p 139 England.
Brent & Michael, R. (2005). Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Research*, 15(12), pp 1777–1786 Cold Spring Harbor Laboratory Press.
Brogna, S. & Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature Structural & Molecular Biology*, 16(2), pp 107–113.
Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S. & Hubbard, T. (2011). Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Research*, 21(5), pp 756–767 Cold Spring Harbor Laboratory Press.
Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P. G. A., Malmstrom, J., Koehler, K., Schrimpf, S., Krijgsveld, J., Kregenow, F., Heck, A. J. R., Hafen, E., Schlapbach, R. & Aebersold, R. (2007). A high-quality catalog of the Drosophila melanogaster proteome. *Nat Biotechnol*, 25(5), pp 576–583 United States.
Bruno, I. G., Karam, R., Huang, L., Bhardwaj, A., Lou, C. H., Shum, E. Y., Song, H.-W., Corbett, M. A., Gifford, W. D., Gecz, J., Pfaff, S. L. & Wilkinson, M. F. (2011). Identification of a MicroRNA that Activates Gene Expression by Repressing Nonsense-Mediated RNA Decay. *Molecular Cell*, 42(4), pp 500–510 Elsevier.
Bryant, D. W., Priest, H. D. & Mockler, T. C. (2012). Detection and Quantification of Alternative Splicing Variants Using RNA-seq. In: *RNA Abundance Analysis*. pp 97–110. Totowa, NJ: Humana Press. (Methods in Molecular Biology). ISBN 978-1-61779-838-2.
Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A. & Babu, M. M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell*, 46(6), pp 871–883.
Bunger, M. K., Cargile, B. J., Sevinsky, J. R., Deyanova, E., Yates, N. A., Hendrickson, R. C. & Stephenson, J. L. (2007). Detection and Validation of Non-synonymous Coding SNPs from Orthogonal Analysis of Shotgun Proteomics Data. *J Proteome Res*, 6(6), pp 2331–2340 American Chemical Society.
Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V. & Briggs, S. P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. *Proceedings of the National Academy of Sciences*, 105(52), pp 21034–21038.
Chang, K.-Y., Georgianna, D. R., Heber, S., Payne, G. A. & Muddiman, D. C. (2010). Detection of Alternative Splice Variants at the Proteome Level in Aspergillus flavus.

Journal of Proteome Research, 9(3), pp 1209–1217.
Cheng, Z. & Menees, T. M. (2011). RNA splicing and debranching viewed through analysis of RNA lariats. *Molecular genetics and genomics*, 286(5-6), pp 395–410 Springer-Verlag.
Cherry, J. M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R. K. & Botstein, D. (1997). Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(6632 Suppl), pp 67–73 ENGLAND.
Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Schölkopf, B., Nordborg, M., Rätsch, G., Ecker, J. R. & Weigel, D. (2007). Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana. *Science*, 317(5836), pp 338–342 American Association for the Advancement of Science.
Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M., Sherry, S., Flicek, P. & Consortium, T. 1. G. P. (2012). The 1000 Genomes Project: data management and community access. *Nat Meth*, 9(5), pp 459–462 Nature Publishing Group.
Consortium, T. E. P. & Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) Nature Publishing Group.
Cox, J. & Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem*, 80, pp 273–299 United States.
Craig, R. & Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics (Oxford, England)*, 20(9), pp 1466–1467 England.
Craig, R., Cortens, J. P. & Beavis, R. C. (2004). Open Source System for Analyzing, Validating, and Storing Protein Identification Data. *Journal of Proteome Research*, 3(6), pp 1234–1242.
David, C. J. & Manley, J. L. (2008). The search for alternative splicing regulators: new approaches offer a path to a splicing code. *Genes & Development*, 22(3), pp 279–285 Cold Spring Harbor Lab.
Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)*, 23(6), pp 673–679 Oxford University Press.
Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N. & Aebersold, R. (2006). The PeptideAtlas project. *Nucleic Acids Research*, 34(suppl 1), pp D655–D658 Oxford University Press.
Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R. & Gingeras, T. R. (2012). Landscape of

transcription in human cells. *Nature*, 489(7414), pp 101–108 Nature Publishing Group.
Eisenhaber, B. & Eisenhaber, F. (2010). Prediction of posttranslational modification of proteins from their amino acid sequence. *Methods in Molecular Biology*, 609, pp 365–384 Springer.
ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*, 9(4), p e1001046 United States.
Eyras, E. (2004). ESTGenes: Alternative Splicing From ESTs in Ensembl. *Genome Res*, 14(5), pp 976–987 Cold Spring Harbor Lab.
Ezkurdia, I., Calvo, E., Del Pozo, A., Vazquez, J., Valencia, A. & Tress, M. L. (2015a). The potential clinical impact of the release of two drafts of the human proteome. *Expert Review of Proteomics*, pp 1–15 Informa Healthcare.
Ezkurdia, I., Del Pozo, A., Frankish, A., Rodriguez, J. M., Harrow, J., Ashman, K., Valencia, A. & Tress, M. L. (2012a). Comparative Proteomics Reveals a Significant Bias Toward Alternative Protein Isoforms with Conserved Structure and Function. *Mol Biol Evol*, 29(9), pp 2265–2283 Oxford University Press.
Ezkurdia, I., Del Pozo, A., Frankish, A., Rodriguez, J. M., Harrow, J., Ashman, K., Valencia, A. & Tress, M. L. (2012b). Comparative proteomics reveals a significant bias towards alternative protein isoforms with conserved structure and function. *Molecular Biology and Evolution*.
Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. & Tress, M. L. (2014a). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23(22), pp 5866–5878 Oxford University Press.
Ezkurdia, I., Rodriguez, J. M., Pau, E. C.-D. S., Vazquez, J., Valencia, A. & Tress, M. L. (2015b). Most Highly Expressed Protein-Coding Genes Have a Single Dominant Isoform. *Journal of proteome ...*, 14(4), pp 1880–1887 American Chemical Society.
Ezkurdia, I., Tress, M. L. & Valencia, A. (2013). Alternative Splicing. In: *Encyclopedia of Biophysics*. pp 48–53. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-16711-9.
Ezkurdia, I., Vazquez, J., Valencia, A. & Tress, M. (2014b). Analyzing the First Drafts of the Human Proteome. *J Proteome Res*, 13(8), pp 3854–3855 American Chemical Society.
Farajollahi, S. & Maas, S. (2010). Molecular diversity through RNA editing: a balancing act. *Trends in Genetics*, 26(5), pp 221–230.
Farrah, T., Deutsch, E. W., Hoopmann, M. R., Hallows, J. L., Sun, Z., Huang, C.-Y. & Moritz, R. L. (2012). The State of the Human Proteome in 2012 as Viewed through PeptideAtlas. *J Proteome Res*, 12(1), pp 162–171 American Chemical Society.
Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., Wong, W. K. & Mockler, T. C. (2009). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research*, 20(1), pp 45–58.
Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. & Bateman, A. (2009). The Pfam protein families database. *Nucleic Acids Research*, 38(Database), pp D211–D222.
Frazer, K. A., Eskin, E., Kang, H. M., Bogue, M. A., Hinds, D. A., Beilharz, E. J., Gupta, R. V., Montgomery, J., Morenzoni, M. M., Nilsen, G. B., Pethiyagoda, C. L., Stuve, L. L., Johnson, F. M., Daly, M. J., Wade, C. M., Cox, D. R., Frazer, K. A., Eskin, E.,

Kang, H. M., Bogue, M. A., Hinds, D. A., Beilharz, E. J., Gupta, R. V., Montgomery, J., Morenzoni, M. M., Nilsen, G. B., Pethiyagoda, C. L., Stuve, L. L., Johnson, F. M., Daly, M. J., Wade, C. M. & Cox, D. R. (2007). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*, 448(7157), p 1050 Nature Publishing Group.

Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., Pozo, A. D., Tress, M., Johnson, R., Guigó, R. & Valencia, A. (2012). Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Research*, 22(7), pp 1231–1242 Cold Spring Harbor Laboratory Press.

Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C. & Yates, J. R. (2000). Automated Identification of Amino Acid Sequence Variations in Proteins by HPLC/Microspray Tandem Mass Spectrometry. *Anal Chem*, 72(4), pp 757–763 American Chemical Society.

Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. (2012). Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol Cell Proteomics*, 11(3), pp M111.014050–M111.014050 American Society for Biochemistry and Molecular Biology.

Gerashchenko, M. V. & Gladyshev, V. N. (2014). Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Research*, 42(17), pp e134–e134 Oxford University Press.

Gilbert, W. (1978). Why genes in pieces? *Nature*, 271(5645), p 501.

Gillingham, A. K., Pfeifer, A. C. & Munro, S. (2002). CASP, the Alternatively Spliced Product of the Gene Encoding the CCAAT-Displacement Protein Transcription Factor, Is a Golgi Membrane Protein Related to Giantin. *Molecular Biology of the Cell*, 13(11), pp 3761–3774 American Society for Cell Biology.

Gnad, F., Gunawardena, J. & Mann, M. (2011). PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Research*, 39(suppl 1), pp D253–D260 Oxford Univ Press.

Gonzàlez-Porta, M., Frankish, A., Rung, J. & Harrow, J. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome …*

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., La Cruz, De, N., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Müller, H.-M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E. M., Tuli, M. A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L. D., Spieth, J. & Sternberg, P. W. (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*, 38(suppl 1), pp D463–D467 Oxford University Press.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.-K., Chrast, J., Lagarde, J., Gilbert, J. G. R., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S. E. & Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome biology*, 7 Suppl 1, pp S4.1–9 England.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein,

M., Guigó, R. & Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9), pp 1760–1774 United States: Cold Spring Harbor Lab.

Haverty, P. M., Weng, Z., Best, N. L., Auerbach, K. R., Hsiao, L.-L., Jensen, R. V. & Gullans, S. R. (2002). HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Research*, 30(1), pp 214–217.

Hawkin, J. D. (1988). A survey on intron and exon lengths. *Nucleic Acids Research*, 16(21), pp 9893–9908 Oxford University Press.

Hayer, K. E., Pizarro, A., Lahens, N. F., Hogenesch, J. B. & Grant, G. R. (2015). Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics (Oxford, England)*, p btv488 Oxford University Press.

He, C., Zhou, F., Zuo, Z., Cheng, H. & Zhou, R. (2009). A Global View of Cancer-Specific Transcript Variants by Subtractive Transcriptome-Wide Analysis.(Bauer, J. A., Ed) *PLoS One*, 4(3), p e4732 Public Library of Science.

Hegyi, H., Kalmar, L., Horvath, T. & Tompa, P. (2011). Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. *Nucleic Acids Research*, 39(4), pp 1208–1219 Oxford University Press.

Hiller, M., Nikolajewa, S., Huse, K., Szafranski, K., Rosenstiel, P., Schuster, S., Backofen, R. & Platzer, M. (2007). TassDB: a database of alternative tandem splice sites. *Nucleic Acids Research*, 35(Database issue), pp D188–192.

Horiuchi, T. & Aigaki, T. (2006). Alternative trans-splicing: a novel mode of pre-mRNA processing. *Biol Cell*, 98(2), pp 135–140 England.

Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E. & Zhang, B. (2004). PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6), pp 1551–1561 Germany.

Hu, Z., Scott, H. S., Qin, G., Zheng, G., Chu, X., Xie, L., Adelson, D. L., Oftedal, B. E., Venugopal, P., Babic, M., Hahn, C. N., Zhang, B., Wang, X., Li, N. & Wei, C. (2015). Revealing Missing Human Protein Isoforms Based on Ab Initio Prediction, RNA-seq and Proteomics. *Scientific reports*, 5, p 10940 Nature Publishing Group.

Huang, D. W., Sherman, B. T., Lempicki, R. A., Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), p 44 Nature Publishing Group.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. & Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research*, 30(1), pp 38–41.

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp 799–816.

Isken, O. & Maquat, L. E. (2008). The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nature Reviews Genetics*, 9(9), pp 699–712.

Iwasaki, R., Kiuchi, H., Ihara, M., Mori, T., Kawakami, M. & Ueda, H. (2009). Trans-splicing as a novel method to rapidly produce antibody fusion proteins. *Biochemical and Biophysical Research Communications*, 384(3), pp 316–321.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R. & Shoemaker, D. D. (2003). Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science*, 302(5653), pp 2141–2144 American Association for the Advancement of Science.

Keren, H., Lev-Maor, G. & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5), pp 345–355 Nature Publishing Group.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H. & Pandey, A. (2014). A draft map of the human proteome. *Nature*, 509(7502), pp 575–581 Nature Publishing Group.

Kim, P., Yoon, S., Kim, N., Lee, S., Ko, M., Lee, H., Kang, H., Kim, J. & Lee, S. (2010). ChimerDB 2.0—a knowledgebase for fusion genes updated. *Nucleic Acids Research*, 38(suppl 1), pp D81–D85 Oxford University Press.

Koenig, T., Menze, B. H., Kirchner, M., Monigatti, F., Parker, K. C., Patterson, T., Steen, J. J., Hamprecht, F. A. & Steen, H. (2008). Robust Prediction of the MASCOT Score for an Improved Quality Assessment in Mass Spectrometric Proteomics. *J Proteome Res*, 7(9), pp 3708–3717 American Chemical Society.

Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S. & Sunyaev, S. (2003a). Increase of functional diversity by alternative splicing. *Trends in Genetics: TIG*, 19(3), pp 124–128.

Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S. & Sunyaev, S. (2003b). Increase of functional diversity by alternative splicing. *Trends in Genetics*, 19(3), pp 124–128 Elsevier.

Lahens, N. F., Kavakli, I. H., Zhang, R. & Hayer, K. (2014). IVT-seq reveals extreme bias in RNA sequencing. *Genome ...*.

Lane, L., Bairoch, A., Beavis, R. C., Deutsch, E. W., Gaudet, P., Lundberg, E. & Omenn, G. S. (2014). Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins. *J Proteome Res*, 13(1), pp 15–20 American Chemical Society.

Lareau, L. F. & Brenner, S. E. (2015). Regulation of Splicing Factors by Alternative Splicing and NMD Is Conserved between Kingdoms Yet Evolutionarily Flexible. *Mol Biol Evol*, 32(4), pp 1072–1079 Oxford University Press.

Legrain, P., Aebersold, R., Archakov, A., Bairoch, A., Bala, K., Beretta, L., Bergeron, J., Borchers, C. H., Corthals, G. L., Costello, C. E., Deutsch, E. W., Domon, B., Hancock, W., He, F., Hochstrasser, D., Marko-Varga, G., Salekdeh, G. H., Sechi, S., Snyder, M., Srivastava, S., Uhlen, M., Wu, C. H., Yamamoto, T., Paik, Y.-K. & Omenn, G. S. (2011). The Human Proteome Project: Current State and Future Direction. *Mol Cell Proteomics*, 10(7), pp M111.009993–M111.009993 American Society for Biochemistry and Molecular Biology.

Levin, Y., Hradetzky, E. & Bahn, S. (2011). Quantification of proteins using data-independent analysis (MSE) in simple andcomplex samples: A systematic evaluation. 11(16), pp 3273–3287 WILEY-VCH Verlag.

Lievens, P. M., Tufarelli, C., Donady, J. J., Stagg, A. & Neufeld, E. J. (1997). CASP, a novel, highly conserved alternative-splicing product of the CDP/cut/cux gene, lacks cut-repeat and homeo DNA-binding domains, and interacts with full-length CDP in vitro. *Gene*, 197(1-2), pp 73–81 NETHERLANDS.

Light, S. & Elofsson, A. (2013). The impact of splicing on protein domain architecture. *Current opinion in structural biology*, 23(3), pp 451–458.

Liu, T. & Lin, K. (2015). The distribution pattern of genetic variation in the transcript isoforms of the alternatively spliced protein-coding genes in the human genome. *Molecular BioSystems*, 11(5), pp 1378–1388 Royal Society of Chemistry.

Long, J. & Caceres, J. (2009). *The SR protein family of splicing factors: master regulators of gene expression*. Biochem. J.

Lopez, G., Maietta, P., Rodriguez, J. M., Valencia, A. & Tress, M. L. (2011). firestar--advances in the prediction of functionally important residues. *Nucleic Acids Research*, 39(Web Server issue), pp W235–41 England.

Low, T. Y., van Heesch, S., van den Toorn, H., Giansanti, P., Cristobal, A., Toonen, P., Schafer, S., Hübner, N., van Breukelen, B., Mohammed, S., Cuppen, E., Heck, A. J. R. & Guryev, V. (2013). Quantitative and Qualitative Proteome Characteristics Extracted from In-Depth Integrated Genomics and Proteomics Analysis. *Cell reports*, 5(5), pp 1469–1478 Elsevier.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett*, 579(9), pp 1900–1903 Netherlands.

Ly, T., Ahmad, Y., Shlien, A., Soroka, D., Mills, A., Emanuele, M. J., Stratton, M. R., Lamond, A. I. & Pines, J. (2014). A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells.(Pines, J., Ed) *eLife*, 3, p e01630 eLife Sciences Publications Limited.

Lykke-Andersen, J. & Bennett, E. J. (2014). Protecting the proteome: Eukaryotic cotranslational quality control pathways. *The Journal of cell biology*, 204(4), pp 467–476 Rockefeller Univ Press.

Maiolica, A., Jünger, M., Ezkurdia, I. & Aebersold, R.Targeted proteome investigation via Selected Reaction Monitoring Mass Spectrometry.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.-F., Revil, T. & Chabot, B. (2007). hnRNP proteins and splicing control. *Advances in Experimental Medicine and Biology*, 623, pp 123–147.

Matlin, A. J., Clark, F. & Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews. Molecular Cell Biology*, 6(5), pp 386–398 Nature Publishing Group.

Maximilian Wei-Lin Popp, L. E. M. (2013). Organizing Principles of Mammalian Nonsense-Mediated mRNA Decay. *Annual review of genetics*, 47(1), pp 139–165 NIH Public Access.

McGlincy, N. J. & Smith, C. W. J. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci*, 33(8), pp 385–393.

McIntyre, L. M., Bono, L. M., Genissel, A., Westerman, R., Junk, D., Telonis-Scott, M., Harshman, L., Wayne, M. L., Kopp, A. & Nuzhdin, S. V. (2006). Sex-specific expression of alternative transcripts in Drosophila. *Genome biology*, 7(8), p R79.

Melamud, E. & Moult, J. (2009). Structural implication of splicing stochastics. *Nucleic Acids Research*, 37(14), pp gkp444–4872 Oxford University Press.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Consortium, T. G., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G. & Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science*, 348(6235), pp 660–665 American Association for the Advancement of Science.

Menon, R. & Omenn, G. S. (2010). Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Research*, 70(9), pp 3440–3449 American Association for Cancer Research.

Menon, R., Zhang, Q., Zhang, Y., Fermin, D., Bardeesy, N., DePinho, R. A., Lu, C., Hanash, S. M., Omenn, G. S. & States, D. J. (2009). Identification of Novel Alternative Splice Isoforms of Circulating Proteins in a Mouse Model of Human Pancreatic Cancer. *Cancer Research*, 69(1), pp 300–309 American Association for Cancer Research.

Meyer, A. J., Almendrala, D. K., Go, M. M. & Krauss, S. W. (2011). Structural protein 4.1R is integrally involved in nuclear envelope protein localization, centrosome–nucleus association and transcriptional signaling. *J Cell Sci*, 124(9), pp 1433–1444 The Company of Biologists Ltd.

Mironov, A. A. (1999). Frequent Alternative Splicing of Human Genes. *Genome Research*, 9(12), pp 1288–1293 Cold Spring Harbor Lab.

Modrek, B. & Lee, C. J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature genetics*, 34(2), pp 177–180 Nature Publishing Group.

Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001a). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, 29(13), pp 2850–2859 Oxford University Press.

Modrek, B., Resch, A., Grasso, C. & Lee, C. (2001b). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, 29(13), pp 2850–2859.

Mollet, I. G., Ben-Dov, C., Felício-Silva, D., Grosso, A. R., Eleutério, P., Alves, R., Staller, R., Silva, T. S. & Carmo-Fonseca, M. (2010). Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. *Nucleic Acids Research*, 38(14), pp 4740–4754 Oxford University Press.

Moore, R. E., Young, M. K. & Lee, T. D. (2002). Qscore: An algorithm for evaluating SEQUEST database search results. *Journal of the American Society for Mass*

Spectrometry, 13(4), pp 378–386.
Mudge, J. M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., Reymond, A., Guigó, R., Hubbard, T. & Harrow, J. (2011). The origins, evolution, and functional potential of alternative splicing in vertebrates. *Molecular Biology and Evolution*, 28(10), pp 2949–2959 United States.
Munoz, J., Low, T. Y., Kok, Y. J., Chin, A., Frese, C. K., Ding, V., Choo, A. & Heck, A. J. R. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. *Mol Syst Biol*, 7(1), pp 550–550 EMBO Press.
Nagai, K., Muto, Y. & Krummel, D. (2001). Structure and assembly of the spliceosomal snRNPs. *Biochemical Society ….*
Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S. & Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology*, 7, p 548 England.
Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nat Meth*, 11(11), pp 1114–1125 Nature Publishing Group.
Nesvizhskii, A. I., Roos, F. F., Grossmann, J., Vogelzang, M., Eddes, J. S., Gruissem, W., Baginsky, S. & Aebersold, R. (2006). Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics*, 5(4), pp 652–670 United States.
Nilsen, T. W. & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280), pp 457–463 England.
Ning, K. & Nesvizhskii, A. (2010). The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics*, 11(Suppl 11), p S14.
Norman, N.Sharpless (2005). INK4a/ARF: A multifunctional tumor suppressor locus. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 576(1-2), pp 22–38.
Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31(13), pp 3635–3641 England.
Omenn, G. S., Lane, L., Lundberg, E. K., Beavis, R. C., Nesvizhskii, A. I. & Deutsch, E. W. (2015). Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J Proteome Res*, 14(9), pp 3452–3460 American Chemical Society.
Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B. B., Simpson, R. J., Eddes, J. S., Kapp, E. A., Moritz, R. L., Chan, D. W., Rai, A. J., Admon, A., Aebersold, R., Eng, J., Hancock, W. S., Hefta, S. A., Meyer, H., Paik, Y.-K., Yoo, J.-S., Ping, P., Pounds, J., Adkins, J., Qian, X., Wang, R., Wasinger, V., Wu, C. Y., Zhao, X., Zeng, R., Archakov, A., Tsugita, A., Beer, I., Pandey, A., Pisano, M., Andrews, P., Tammen, H., Speicher, D. W. & Hanash, S. M. (2006). *Overview of the HUPO Plasma Proteome Project: Results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database*. *Exploring the Human Plasma Proteome* pp 1–35 Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. (OMENN:HUMAN PLASMA PROTEO O-BK). ISBN 9783527609482.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12), pp 1413–1415 United States.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., Babak, T., Siu, H., Hughes, T. R., Morris, Q. D., Frey, B. J. & Blencowe, B. J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Molecular Cell*, 16(6), pp 929–941.

Penny, D., Hoeppner, M. P., Poole, A. M. & Jeffares, D. C. (2009). An Overview of the Introns-First Theory. *Journal of Molecular Evolution*, 69(5), pp 527–540 Springer-Verlag.

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K. & White, O. (2001). The Comprehensive Microbial Resource. *Nucleic Acids Research*, 29(1), pp 123–125 Oxford University Press.

Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. (2010). Noisy Splicing Drives mRNA Isoform Diversity in Human Cells.(Dermitzakis, E. T., Ed) *PLoS Genet*, 6(12), p e1001236 Public Library of Science.

Pirrotta, V. (2002). Trans-splicing in Drosophila. *Bioessays*, 24(11), pp 988–991 Wiley Subscription Services, Inc., A Wiley Company.

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., Hart, E., Suner, M.-M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., DiCuccio, M., Kellis, M., Lee, J., Lin, M. F., Schuster, M., Shkeda, A., Amid, C., Brown, G., Dukhanina, O., Frankish, A., Hart, J., Maidak, B. L., Mudge, J., Murphy, M. R., Murphy, T., Rajan, J., Rajput, B., Riddick, L. D., Snow, C., Steward, C., Webb, D., Weber, J. A., Wilming, L., Wu, W., Birney, E., Haussler, D., Hubbard, T., Ostell, J., Durbin, R. & Lipman, D. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, 19(7), pp 1316–1323 Cold Spring Harbor Lab.

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., Goebl, M. G. & Iakoucheva, L. M. (2010). Identification, analysis, and prediction of protein ubiquitination sites. *Proteins: Structure, Function, and Bioinformatics*, 78(2), pp 365–380 Wiley Online Library.

Raitano, A. B., Halpern, J. R., Hambuch, T. M. & Sawyers, C. L. (1995). The Bcr-Abl leukemia oncogene activates Jun kinase and requires Jun for transformation. *Proceedings of the National Academy of Sciences*, 92(25), pp 11746–11750 National Acad Sciences.

Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O. & Aebersold, R. (2009). Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol Cell Proteomics*, 8(11), pp 2405–2417 American Society for Biochemistry and Molecular Biology.

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A. & Tress, M. L. (2012). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research*.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2), pp 85–94 Oxford University Press.

Ryu, G.-M., Song, P., Kim, K.-W., Oh, K.-S., Park, K.-J. & Kim, J. H. (2009). Genome-

wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Research*, 37(4), pp 1297–1307.
S E Leff, A. & Rosenfeld, M. G. (2003). Complex Transcriptional Units: Diversity in Gene Expression by Alternative RNA Processing. *dx.doi.org*, 55(1), pp 1091–1117 Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA.
Sammeth, M., Foissac, S. & Guigó, R. (2008). A general definition and nomenclature for alternative splicing events. *PLoS computational biology*, 4(8), p e1000147 United States.
Savitski, M. M., Wilhelm, M., Hahne, H., Küster, B. & Bantscheff, M. (2015). A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol Cell Proteomics*, 14(9), pp mcp.M114.046995–2404 American Society for Biochemistry and Molecular Biology.
Schacherer, J., Shapiro, J. A., Ruderfer, D. M., Kruglyak, L., Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature*, 458(7236), p 342 Nature Publishing Group.
Schubert, U., Antón, L. C., Gibbs, J., Norbury, C. C., Yewdell, J. W. & Bennink, J. R. (2000). Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature*, 404(6779), pp 770–774 Nature Publishing Group.
Severing, E. I., van Dijk, A. D. J., Stiekema, W. J. & van Ham, R. C. H. J. (2009a). Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics*, 10, p 154.
Severing, E. I., van Dijk, A. D., Stiekema, W. J. & van Ham, R. C. (2009b). Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome. *BMC Genomics*, 10(1), p 154 BioMed Central Ltd.
Shepard, P. J. & Hertel, K. J. (2008). Conserved RNA secondary structures promote alternative splicing. *RNA*, 14(8), pp 1463–1469 Cold Spring Harbor Lab.
Sheynkman, G. M., Shortreed, M. R., Frey, B. L. & Smith, L. M. (2013). Discovery and Mass Spectrometric Analysis of Novel Splice-junction Peptides Using RNA-Seq. *Mol Cell Proteomics*, 12(8), pp 2341–2353 American Society for Biochemistry and Molecular Biology.
Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. (2013). Combining Results of Multiple Search Engines in Proteomics. *Mol Cell Proteomics*, 12(9), pp 2383–2393 American Society for Biochemistry and Molecular Biology.
Silva, D., Clapham, P., Coates, G., Fairley, S. & Fitzgerald, S. (2013). *Ensembl 2013*. Nucleic Acids Res.
Simon, D. N. & Wilson, K. L. (2011). The nucleoskeleton as a genome-associated dynamic "network of networks." *Nature Reviews. Molecular Cell Biology*, 12(11), pp 695–708 Nature Publishing Group.
Smith, C. W. & Valcárcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci*, 25(8), pp 381–388 ENGLAND.
Sorek, R., Shamir, R. & Ast, G. (2004). How prevalent is functional alternative splicing in the human genome? *Trends in Genetics*, 20(2), pp 68–71.
Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Consortium, T. R., Hubbard, T. J., Guigó, R., Harrow, J. & Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat Meth*, 10(12), pp 1177–1184 Nature Publishing Group.
Sugnet, C. W., Srinivasan, K., Clark, T. A., O'Brien, G., Cline, M. S., Wang, H., Williams,

A., Kulp, D., Blume, J. E., Haussler, D. & Ares, M. J. (2006). Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS computational biology*, 2(1), p e4.

Taneri, B., Snyder, B. & Gaasterland, T. (2011). Distribution of Alternatively Spliced Transcript Isoforms within Human and Mouse Transcriptomes. *Journal of OMICS Research*, 1(1), pp 1–5.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S. P. & Bafna, V. (2007a). Improving gene annotation using peptide mass spectrometry. *Genome Research*, 17(2), pp 231–239.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S. P. & Bafna, V. (2007b). Improving gene annotation using peptide mass spectrometry. *Genome Res*, 17(2), pp 231–239 Cold Spring Harbor Lab.

Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J. & Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology*, 16(9), pp 996–1001.

Tress, M. L., Bodenmiller, B., Aebersold, R. & Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome biology*, 9(11), p R162 England: BioMed Central Ltd.

Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. A., Wesselink, J. J., Yeats, C., Ólason, P. Í., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R. A., López, G., Sadowski, M. I., Watson, J. D., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Størling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramírez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S. E., Reymond, A., Birney, E., Brunak, S., Casadio, R., Guigo, R., Harrow, J., Hermjakob, H., Jones, D. T., Lengauer, T., A Orengo, C., Patthy, L., Thornton, J. M., Tramontano, A. & Valencia, A. (2007a). The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences*, 104(13), pp 5495–5500.

Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. A., Wesselink, J. J., Yeats, C., Ólason, P. Í., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R. A., López, G., Sadowski, M. I., Watson, J. D., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Størling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramírez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S. E., Reymond, A., Birney, E., Brunak, S., Casadio, R., Guigó, R., Harrow, J., Hermjakob, H., Jones, D. T., Lengauer, T., Orengo, C. A., Patthy, L., Thornton, J. M., Tramontano, A. & Valencia, A. (2007b). The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences*, 104(13), pp 5495–5500 National Acad Sciences.

Tress, M., Cheng, J., Baldi, P., Joo, K., Lee, J., Seo, J.-H., Lee, J., Baker, D., Chivian, D., Kim, D. & Ezkurdia, I. (2007c). Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins*, 69 Suppl 8, pp 137–151 United States.

Trinh, C. H., Asipu, A., Bonthron, D. T. & Phillips, S. E. V. (2009). Structures of alternatively spliced isoforms of human ketohexokinase. *Acta Crystallogr D Biol Crystallogr*, 65(Pt 3), pp 201–211 Denmark.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., Zhang, H. & Consortium, T. F. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids*

Research, 37(suppl 1), pp D555–D559 Oxford University Press.
Uhlen, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., Feilitzen, von, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Heijne, von, G., Nielsen, J. & Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220), pp 1260419–1260419 American Association for the Advancement of Science.
Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J. & Darnell, R. B. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119), pp 580–586.
Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J.-S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., Zeeberg, B. R., Kane, D., Weinstein, J. N., Blume, J. & Darnell, R. B. (2005). Nova regulates brain-specific splicing to shape the synapse. *Nat Genet*, 37(8), pp 844–852.
Vanderperre, B., Lucier, J.-F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.-M. & Roucou, X. (2013). Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome.(Jan, E., Ed) *PLoS One*, 8(8), p e70698 Public Library of Science.
Venables, J. P., Koh, C.-S., Froehlich, U., Lapointe, E., Couture, S., Inkel, L., Bramard, A., Paquet, E. R., Watier, V., Durand, M., Lucier, J.-F., Gervais-Bird, J., Tremblay, K., Prinos, P., Klinck, R., Elela, S. A. & Chabot, B. (2008). Multiple and specific mRNA processing targets for the major human hnRNP proteins. *Molecular and Cellular Biology*, 28(19), pp 6033–6043.
Vogel, C. & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4), pp 227–232 Nature Publishing Group.
Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), pp 470–476.
Wang, J., Li, S., Zhang, Y., Zheng, H., Xu, Z., Ye, J., Yu, J., Wong, G. K.-S., Wang, J., Li, S., Zhang, Y., Zheng, H., Xu, Z., Ye, J., Yu, J. & Wong, G. K.-S. (2003). Vertebrate gene predictions and the problem of large genes. *Nature Reviews Genetics*, 4(9), p 741 Nature Publishing Group.
Wang, Z. & Burge, C. B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5), pp 802–813 Cold Spring Harbor Lab.
Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Küster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502), pp 582–587 Nature Publishing Group.
Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F. & Williams, K. L. (1996). Progress with proteome projects: why all

proteins expressed by a genome should be identified and how to do it. *Biotechnology & genetic engineering reviews*, 13, p 19.
Will, C. L. & Lührmann, R. (2011). Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology*, 3(7), pp a003707–a003707 Cold Spring Harbor Lab.
Wu, Q. & Krainer, A. R. (1999). AT-AC Pre-mRNA Splicing Mechanisms and Conservation of Minor Introns in Voltage-Gated Ion Channel Genes. *Molecular and Cellular Biology*, 19(5), pp 3225–3236 American Society for Microbiology.
Xu, Q., Modrek, B. & Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, 30(17), pp 3754–3766 Oxford University Press.
Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D. L., Sun, H., Fu, X.-D. & Zhang, Y. (2009). Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. *Molecular Cell*, 36(6), pp 996–1006.
Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C. B. (2005). Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences*, 102(8), pp 2850–2855 National Acad Sciences.
Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3(9), pp 698–709 Nature Publishing Group.
Zhang, M. Q. & Zhang, M. Q. (2002). Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3(9), p 698 Nature Publishing Group.